NTNU

Norwegian University of
Science and Technology

Sentiment Analysis in Norwegian Political News

Employing Replicated Methods and
Experimental Features to Understand a
Complex Domain

Patrik Fridberg Bakken
Terje Bratlie

Master of Science in Informatics
Submission date: May 2016
Supervisor: Jon Atle Gulla, IDI

Norwegian University of Science and Technology
Department of Computer and Information Science

Abstract

This thesis employs machine learning in an effort to develop a sentiment analysis engine for the Norwegian political news domain. In combination with computational linguistics and statistics we set out to gain knowledge and understanding of a less researched area of sentiment analysis, which is more complex than other well-known domains. As the mass media is setting the agenda for what should be focused on by the general public, the news world has significant influence on what is subsequently expressed on social media. The motivation for choosing this domain is the lack of research and the fact that if Twitter and Facebook are deemed important platforms for sentiment analysis, the news should be as well.

Replicating proven methods from other well-structured and understood domains, we try to achieve similar precision results in spite of the lack of resources available in the Norwegian language. Evaluating the results from this work led to the discovery of essential characteristics of the political news domain. These characteristics portray the challenges to overcome in order to achieve state-of-the-art classification results. We uncovered that the language in the domain in question is unstructured, sentiment is conveyed in a subtle manner without the use of explicit sentiment-bearing words, and require contextual knowledge.

Further, we experimented with a two-step binary classification method to pinpoint the areas of effect for each feature included in the sentiment engine. Observing the results of each classification step, we note that negation count does not in fact improve performance. However, the exclusion of neutral co-occurring terms in the polarity classification step achieved close to state-of-the-art precision scores. In addition to this, we find that the most imperative area of focus should be on the subjectivity classification step, as improvements here will eventually show a momentous increase in overall precision of the sentiment engine.

Denne avhandlingen benytter maskinlæring i et forsøk på å utvikle en motor for sentimentanalyse i det norske politiske nyhetsdomenet. I kombinasjon med datalingvistikk og statistikk gjorde vi et forsøk på å tilegne oss kunnskap og forståelse om et område med mindre forskningsfokus som er mer komplisert enn andre kjente domener. Siden massemedia setter dagsorden for hva som bør fokuseres på av allmennheten, har nyhetsverden betydelig innflytelse på det som senere blir uttrykt på sosiale medier. Motivasjonen for å velge dette domenet er mangel på forskning og det faktum at hvis Twitter og Facebook anses som viktige plattformer for sentiment analyse, bør nyheter også være det.

Replikasjon av tidligere utprøvde metoder fra andre velstrukturerte og forståtte domener, prøver vi å oppnå samme presisjonsresultater på tross av mangel på ressurser i det norske språket. Evaluering av resultatene fra dette arbeidet førte til oppdagelsen av viktige kjennetegn ved det politiske nyhetesdomenet. Disse egenskapene beskriver utfordringer vi må overvinne for å oppnå "state-of-the-art" klassifiseringsresultater. Vi avdekket at språket i det aktuelle domenet er ustrukturert, meninger formidles på en subtil måte uten bruk av eksplisitte sentimentbærende ord, og krever kontekstuell kunnskap.

Videre har vi eksperimentert med en to-trinns binær klassifiseringsmetode for å finne de områdene der hver funksjon inkludert i sentimentmotoren har mest innvirkning. Observasjon av resultatene fra hvert klassifiseringstrinn, viser at negasjonsantallet ikke faktisk forbedrer ytelsen. Imidlertid oppnådde vi ved ekskludering av nøytrale "co-occurring terms" i polaritetsklassifiseringen nær "state-of-the-art" presisjonsresultater. I tillegg til dette finner vi ut at den mest avgjørende delen å fokusere på bør være subjektivitetsklassifiseringen hvor forbedringer til slutt vil vise en betydningsfull økning i total presisjon av sentimentmotoren.

Preface

This thesis is submitted to the Norwegian University of Science and Technology (NTNU), as partial fulfillment of the degree Master of Computer Science, and as part of the course IT3901 − Informatics Postgraduate Thesis: Software. The work culminating in this report has been performed at the Department of Computer and Information Science (IDI), under the supervision of Professor Jon Atle Gulla, and as a part of the SmartMedia project.

[This page is intentionally left blank.]

Acknowledgements

First and foremost, we would like to direct our thanks to our supervisor, Professor Jon Atle Gulla, for continiously having meetings and brain-storming sessions with us, and always encouraging us to try new things, resulting in the work shown in this thesis. Our thanks should also go to Dr. Jon Espen Ingvaldens for providing us with in-depth understanding of the SmartMedia project and its structure. Additionally we are grateful for data provided by Chief editor of VG, Arnstein Johansen, as well as Trond Johansen and Thomas Oldervoll of NRK. We would like to note if errors occur within this work, they are our own.

[This page is intentionally left blank.]

Contents

List of Figures

[This page is intentionally left blank.]

List of Tables

[This page is intentionally left blank.]

Part I

Research Overview and Summary

[This page is intentionally left blank.]

Chapter 1

Introduction

This chapter introduces the research conducted within the scope of this thesis. Section 1.1 elaborates on the background and motivation for the work presented. In section 1.2 a wider understanding of the problem faced for conducting this research can be found. Research goals and questions, as well as research contributions can be found in section 1.3 and 1.4 respectively. An overview of two papers that have been written and awaiting publication, can be found in section 1.5. Finally, section 1.6 gives a structured overview of the entire thesis.

1.1 Background and Motivation

To come to terms with just how big the World Wide Web (Internet) is and the amount of information it makes readily available, we can take a look at research conducted in 1999 by Lawrence and Giles [2], Albert et al. [3], and Huberman et al. [4]. The Internet was then believed to consist of upwards of 800 million documents on the searchable web. By 2009 it had grown to about 500 exabytes[1] [5], and in 2014, the search engine Google[2] had over 4 million requests every single minute [6], backing up the claim of exponential growth by Huberman et al.

With a focus on news, there were in January of 2015 alone, 780 million unique visitors to the top ten most visited English news site networks [7], and similarly half of the American adult population went online in 2010 to get involved in the midterm elections in one way or another [8]. This demonstrates just how big of a user market there is to exploit, especially as all the information available, and the need for accessing the right kind, is something that, without some structure, is near impossible.

[1] $1 * 10^{18}$ bytes

[2] http://www.google.com/about/

One field trying to make sense of much of this information, is sentiment analysis (often referred to as opinion mining). Its main usage area is to classify span of text as positive, negative, or on a spectrum in-between [9, 10]. According to Pang and Lee [10], most work related to sentiment analysis has had its focus in easier-to-classify text, such as product reviews, where a more structured opinion on a matter can be extracted. One research area that has been of little focus is the news domain, and especially political news. Evgenia and van Der Groot [11] analyzed the bias across languages in news headlines, and similarly bias in media was analyzed by Blaz et al. [12].

Mass media setting the agenda for what should be focused on by the general public is a well-researched area [13, 14, 15, 16]. By setting the agenda, and the key areas to be focused on, say a politician included in a scandal, can shift public opinion for that specific politician, and subsequently is something that can have an impact on which party the electoral vote for come election day. Continuously being fed information, and frequently seeing a certain topic, readers learn how much importance to attach to it. This implies that the news world has significant influence on what is expressed on social media. As Feldman [17] notes, sentiment analysis on Twitter and Facebook can provide substantial information for a politician and their network for understanding how voters feel about certain matters. Hence, if sentiment analysis of expressions posted by the masses is of focus, then political news articles, too, deserve attention. By successfully employing sentiment analysis in the political news domain, the news image would be more complete and transparent, as possible biases in different news sources can be uncovered. This leads to readers forming informed opinions about politicians and political parties with scrutiny, instead of being "puppets of the media". The lack of research in this field, as well as the potential applications, is what has fueled motivation for further work in this thesis.

Most research conducted in the field of sentiment analysis has been done in the English language [9, 10, 18, 19, 20, 21, 22, 23], whereas far less research has been completed in the Norwegian domain. Some examples in the Norwegian domain alone, are work done by Hammer et al. [24, 25], Bai et al. [26], and Njølstad et al. [27]. One reason for it being a language with less research, is the size of the language itself, with it being spoken by only a fraction of English speakers [28]. To further fill the gap in the research community, Norwegian is therefore the language of study in this thesis.

In addition to contributing to the field of sentiment analysis, this thesis is part of work that will be embodied in a larger research project at the Norwegian University of Science and Technology (NTNU). The research project, named SmartMedia[3], with contributions such as [1, 29, 30], has a focus on news recom-

[3]https://www.ntnu.no/wiki/display/smartmedia/SmartMedia+Program

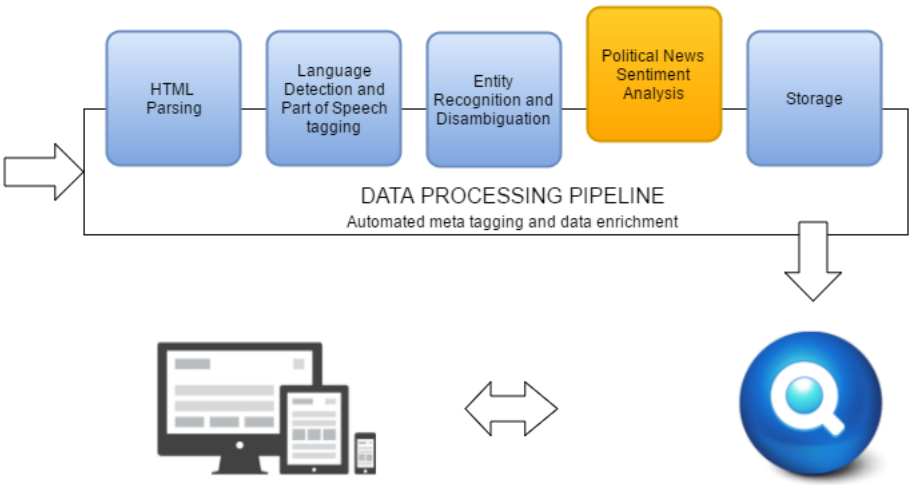


Figure 1.1.1: High-level overview of how our thesis fits into the SmartMedia project. Figure adopted from [1]

mendation, with an additional focus on text analytics and sentiment analysis. As seen from Figure 1.1.1, in the pipeline of data processing, the research conducted in this thesis will be the groundwork for a module (highlighted in yellow) solely focusing on sentiment analysis of Norwegian political text.

1.2 Problem Outline

This thesis has an emphasis on detection of author sentiment in Norwegian news articles collected from two of the biggest news sources in Norway; NRK[4] and VG[5]. As mentioned, this is called sentiment analysis, and even though it is still a young field of research with several challenges yet to overcome, the research that has been done has already shown how powerful sentiment analysis can be. It has been adopted into the commercial market, especially in the product and service review domain where Facebook and Twitter are the most prominent focal points [17]. The financial and political markets are not far behind, although these domains are more complicated and require more advanced techniques. Our biggest problems are the relatively small amount of research done in the domains of political news and the Norwegian language. The challenges introduced by this fact are detailed in sections 1.2.1 and 1.2.2 respectively.

[4]www.nrk.no/nyheter
[5]http://www.vg.no/nyheter/innenriks/norsk-politikk/

1.2.1 News Domain

As Facebook and Twitter are two of the biggest platforms for commercial sentiment analysis products. Much research has been done on the social media platforms as the texts found there are usually well-suited for sentiment analysis applications. The subjective nature of the tweets and Facebook statuses, coupled with a clearly defined and consistent sentiment target makes it easier to build accurate sentiment engines for these platforms. The same is true in the domains of product and service review domains, which can be found on the mentioned platforms.

In the news domain, however, this is not the case. Most categories of news articles are more unstructured. They often include several sentiment targets in the same article − even in the same paragraph − and many news categories can seem very objective. There are also challenges pertaining to the attribute differences between articles in the news domain, and within some news categories as well. Commonly used attributes in sentiment analysis such as text length, use of quotes (with or without quotation marks), use of sarcasm and irony, are just a few of the differences [31].

Each category often have its own jargon which can make it difficult for a standard language-wide sentiment lexicon to catch all the sentiment-bearing expressions in the article. As Njølstad and Høysæter explains in their master's thesis [31], "bull" is a negative sentiment word in the financial domain, but in a standard lexicon it is just the name of an animal. In the political domain, an example in Norwegian is the use of "nyttig idiot" (*"useful idiot"*), which here means a person who is manipulated to promote a cause on another person or group's behalf, without the knowledge of this happening. In other categories and domains, this expression is non-existing. To solve these challenges a sentiment engine needs to be domain-specific enough to catch these sentiment expressions, and be tailored to the other differences found in the domain.

Political news specifically may be one of the more challenging news categories. There are several sentiment targets in most articles, which are unstructured and varies greatly, and the language style can often be misleadingly objective. Take the following excerpt from a Norwegian news article:

> Høyre, Frp, Venstre og delvis KrF retter kritikk mot **Stoltenberg-regjeringens** arbeid med beredskap både før og etter 22. juli. **Grete Faremo (Ap)**, som var justisminister fra 2011 til 2013, får kritikk for at hun ikke fulgte opp beredskapsarbeidet godt nok. "Alle tiltak som er iverksatt for å rette opp disse feilene, er iverksatt av justisminister **Anders Anundsen**", skriver flertallet.

where highlighted text depicts targets. In addition, from the following paragraph, the misleadingly objective language often occurring in political news articles can

be observed:

> Hun sier de valgte å ikke gå ut med avtalen straks den var signert, fordi de ville la det gå litt tid. Det samme kan ikke sies om det nye trekløveret, mener hun.
> *She says they chose to not go public with the agreement as soon as it was signed, because they wanted to let some time pass. The same can not be said about the new trio, she says.*

1.2.2 Language of Analysis

In Section 1.1 we noted that most research in the field of sentiment analysis is done in the English language. Some have been conducted in Norwegian, but it is relatively very little. As Njølstad and Høysæter mentions [31], you cannot do sentiment analysis effectively without extensive knowledge about the language of study. The lack of research in the Norwegian language results in a limited set of lexical and linguistic tools to help build effective sentiment analysis engines. In English, there are publicly available resources; sentiment lexica to look up sentiment values of encountered words, word graphs to assign values to new words, part-of-speech taggers to classify words, to name a few. In Norwegian, some of these tools need to be created from scratch, especially if we want domain-specific lexica. These are essential in sentiment analysis as most researchers believe domain-independent lexica to be greatly ineffective compared to domain-dependent lexica [19, 23]. To put the inferiority of universal lexica into numbers, Wilson et al. [32] achieved a $\sim$17% gain in precision when using a domain-dependent lexicon.

Feldman [17] describes three ways to acquire a domain-specific lexicon when none already exists:

1. the manual approach; coding of the lexicon,
2. the dictionary-based approach; utilizing resources like WordNet[6] to expand a set of seed words,
3. and the corpus-based approach; using a large corpus from a specific domain to expand a set of seed words.

The only approach which is not dependent on having access to lexical tools, such as sentiment lexica or synonym dictionaries et cetera, or a very large dataset of documents in the same domain, is the first one. This manual approach have previously been deemed too extensive by Feldman [17], however, a recent study

[6] https://wordnet.princeton.edu

by Njølstad et al. [33] shows promising results in the financial news domain after acquiring a sentiment lexicon with a reasonable amount of manual labor.

As we are in the situation where we do not have any resources available to us to use any of the automated approaches, we are curious about how well the methods of Njølstad et al. can perform in our domain.

1.3 Research Goals and Questions

The research conducted in this thesis has its main focus on building a sentiment analysis system for Norwegian political news articles. This requires us to 1) analyze and understand the characteristics of political sentiment, 2) analyze the results compared to that of well understood domains, and 3) how this can be improved with domain-specific approaches. The research goals, and in turn the research questions, will now be introduced in turn.

1.3.1 Political News Characteristics

In order for a sentiment analysis system to be employed in a domain where previous research is lacking, one important aspect to address is the news articles themselves. We need to analyze the structure in such articles, how the sentiment is conveyed, and what impact this have on the analysis of its sentiment.

> **RQ1** What characterizes political news and how do these characteristics affect the analysis of their sentiments?

1.3.2 Sentiment Engine Construction

For us to construct a sentiment analysis engine, which at the same time will achieve similar precision scores to that of other domains, a few key areas have to be addressed: we need to create a domain-specific lexicon for the Norwegian political domain, make use of already-existing classification methods, and analyze to see if similar results to that of other domains can been achieved (RQ2). In case of lacklustre results from the mere replication of approaches, we need to further analyze our domain to find features better describing our domain, and employ this knowledge in the implementation of more domain-specific approaches (RQ3).

> **RQ2** To what extent can standard sentiment analysis methods from well understood domains be applied to Norwegian political news?

> **RQ3** How can we improve the analysis of political news sentiments with domain-specific approaches and techniques?

1.4 Research Contributions

This thesis has three main contributions, each in relation to the aforementioned research questions, and are as follows:

> **C1** Political news can be characterized as difficult to extract sentiment from due to the seemingly objective, complex language used, conveying sentiment in subtle ways, thus resulting in many news article paragraphs receiving the wrong classification.
>
> **C2** Standard sentiment analysis methods replicated from well understood domains does not achieve similar results in the Norwegian political news domain, in terms of precision. Analysis suggest that adaptation of methods to the domain in question should be implemented.

As can be observed from Paper I (Chapter 6), a large majority of paragraphs was classified as being objective (neutral). A large amount of the sentiment bearing paragraphs that was incorrectly classified was classified as objective. This amount was significantly larger than the amount of paragraphs incorrectly classified as the opposite polarity. This defends the hypothesis that the political domain is a much more complex one, where sentiment is, if present, conveyed in a much more subtle way than in domains such as product reviews. It also led to the characteristics presented in this thesis, outlining the effect these have on the analysis of the political news sentiment. Subsequently, as our research show, the mere replication of approaches to that of different, more well understood domains, does in fact not have the same impact in the political news domain.

> **C3** A two-step binary classifier can pinpoint the areas of effect of each feature. It also illustrates which areas to focus on and improve to optimize a sentiment engine. The subjectivity classification step lacks the most performance and will increase overall precision the most. The polarity classification step yields close to state-of-the-art precision when excluding neutral Co-Occurring Terms (COTs) as a feature.

From the results achieved in Paper I, we derived the suggestion to implement different approaches, where a main focus were put on subjective paragraphs, in isolated steps. After initial analysis of the paragraphs and characteristics, observations were made that negations occurred more frequently in paragraphs with a negative sentiment. This led to the inclusion of a simple feature, focusing only on the negation count averaged over paragraphs with a different sentiment classification scheme. However, the inclusion of negation count turned out to only

have a marginal decrease in precision in both classification steps, suggesting the difference in negations per paragraph between the two classes in each step was not big enough to make a positive impact. Observing the classifiers employed in the sentiment engine, neutral COTs had an unwanted impact on the decision making with each machine learning classifier in the polarity classification step. As this step only classified sentiment bearing paragraphs, the exclusion of neutral COTs was then experimented with, resulting in better scores than with them present.

1.5 Papers

There are two papers included in this thesis. The first paper is submitted to, and is awaiting notification of acceptance for the 4th International Conference on Statistical Language and Speech Processing (SLSP 2016). The second paper is to be submitted to either the 16th IEEE International Conference on Data Mining (ICDM 2016), or the 26th International Conference on Computational Linguistics (COLING 2016).

Paper I. Patrik F. Bakken, Terje A. Bratlie, and Jon Atle Gulla: *On the Challenges of Political News Sentiment Analysis*, submitted to the 4th International Conference on Statistical Language and Speech Processing (SLSP 2016).

Paper II. Patrik F. Bakken, Terje A. Bratlie, and Jon Atle Gulla: *Understanding the Political News Domain - Analyzing Negation Count and Co-Occurring Terms in Sentiment Analysis*, to be submitted to either the 16th IEEE International Conference on Data Mining (ICDM 2016), or the 26th International Conference on Computational Linguistics (COLING 2016).

We will now give a brief overview of how the two papers in this thesis relate, before giving a short summary of each one. These papers can both be read in full in Part II for the ones interested. As seen from Figure 1.5.1, what is the finding of Paper I, is then used as the purpose for Paper II, interchanging these nicely. Both papers focus on creating a sentiment analysis engine, one that is used for classifying Norwegian political news paragraphs.

In the first paper (Paper I), we have a focus on imitating methods from past studies, to see if similar results can be achieved when analyzing Norwegian political news articles. Additionally, we want to characterize the difficulties of the domain we perform our study on. From Figure 1.5.2, an input to the system consisting of annotated articles (each paragraph with its own annotation) is given. Subsequently, we generate candidate Co-Occurring Terms (COTs) from said dataset,

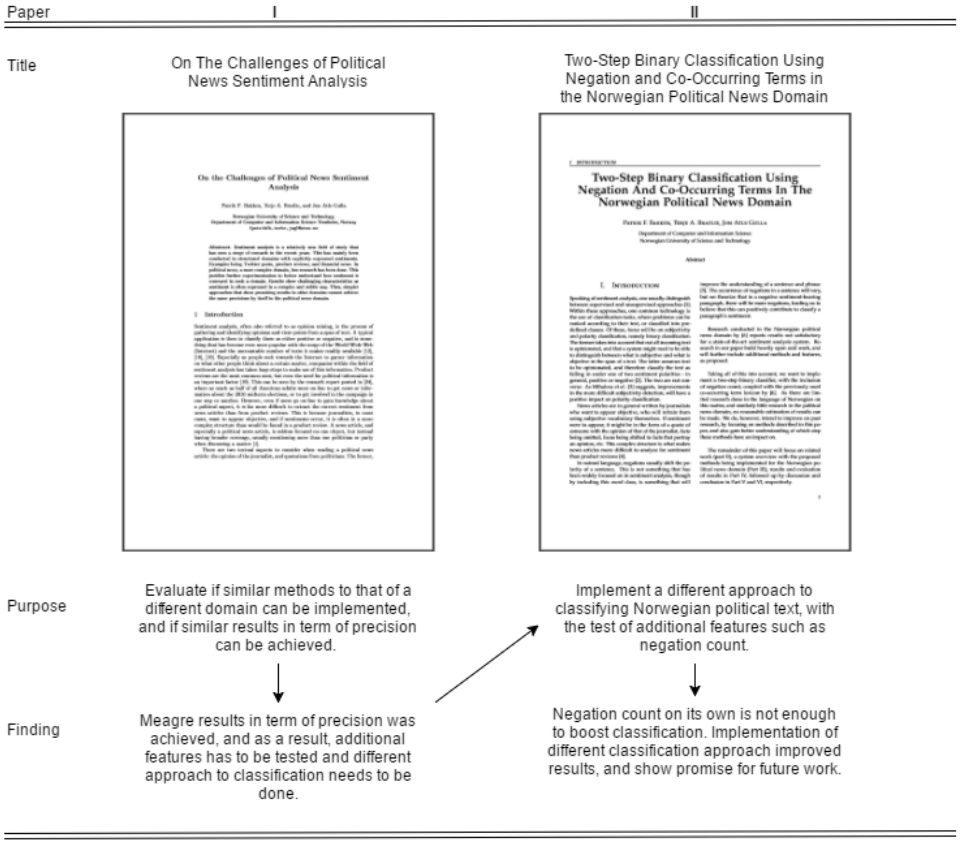


Figure 1.5.1: Overview of how the papers included in this thesis relate.

before ranking these using implemented ranking functions. Top ranked COTs are then annotated, comprising our lexicon which in turn is used for creation of feature vectors used by machine learning classifiers. With the implementation of this system, we find that results to a more structured domain like the financial news domain, can not be replicated in terms of precision scores. We theorize why this is a difficult domain, mainly down to the language of the articles, with complex sentimental structures, conveyed in subtle ways, and characterize a few key areas that should be focused on for future work.

Our second paper (Paper II), tries to tackle some of the limitations put forth from the study of our first paper. We analyze news articles and theorize that with the implementation of an additional feature, a simple negation count per paragraph, will in turn help improve results. Additionally, after results from our first paper, we recognized that a larger majority of the paragraphs being annotated as neutral. Hence, we implement a two-step binary classifier, which will filter out neutral paragraphs and only focus on classifying its polarity in a second step. Fig-

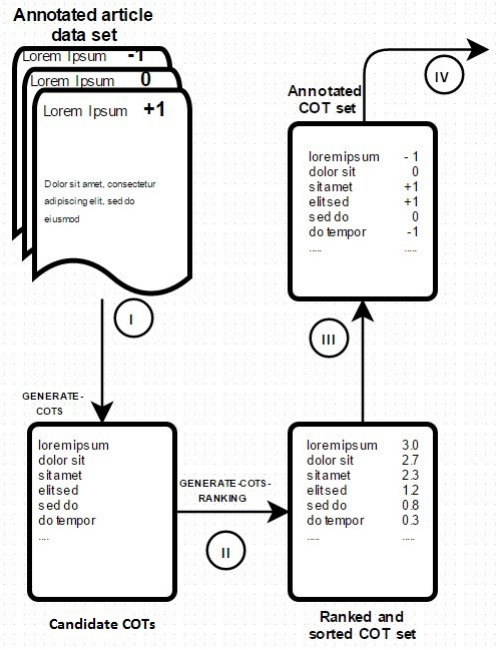


Figure 1.5.2: High-level system overview of sentiment engine from Paper I

ure 1.5.3 depicts a high-level system overview of how the two-step binary classifier operates. First, by having the same annotated data set as from our first paper, we distinguish between paragraphs with polarity and objectivity. Subsequently, we train and test the classifiers, evaluating how well they did. In the second step, only data with polarity is used, still sampled from the annotated data set. This is then passed on for extracting features, given as input to training and testing of classifiers, followed by evaluation. Our findings indicate that with the implementation of a two-step binary classifier, we improve results. However, this is not due the inclusion of negation count, leading us to believe that other combinations of features should be included.

1.6 Thesis Structure

This thesis is made up of two parts. First, we give a background and present relevant information to the research being conducted in this thesis. Within the same part (Chapter 2), a detailed description of used methods and technologies for this work, can be found. The most relevant work related to this thesis are presented in Chapter 3, results and evaluation of results in Chapter 4, before a conclusion on the research and discussion on how to further the research, in Chapter 5. Part

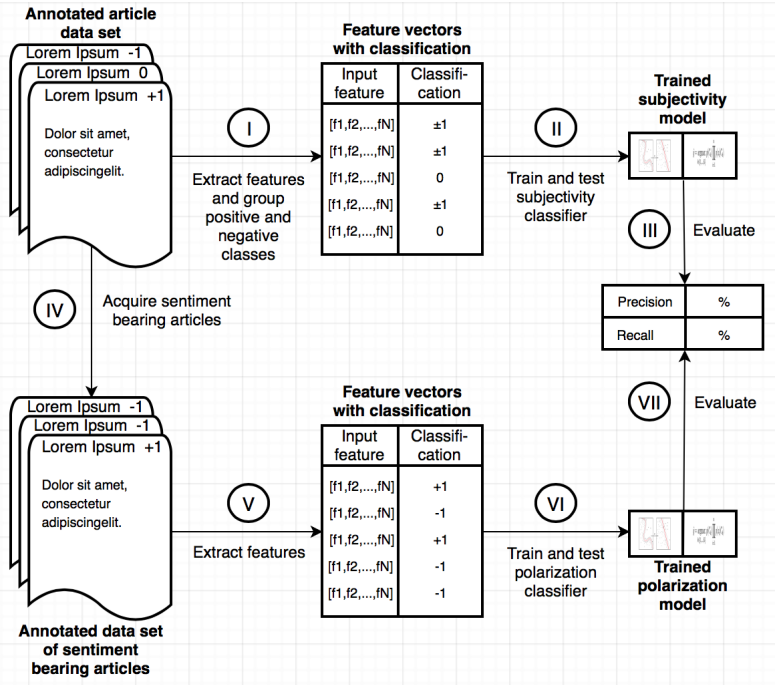


Figure 1.5.3: System overview of two-step binary classification process from Paper II

II focuses on two papers, written and submitted for publication, as is detailed in 1.5.

Part I - Research Overview and Summary

Chapter 1 - Introduction

This chapter describes the background and motivation for this thesis, elaborates on the difficulties within the domain chosen, presents research goals and contributions, as well as an overview of papers written and submitted for publication.

Chapter 2 - Theoretical Background

This chapter explains central methods and terms used for work in this thesis. This includes machine learning algorithms, and computational linguistics.

Chapter 3 - Related Work

This chapter gives a description of selected related work to this thesis. In order to not simply repeat what has been written in the two papers, a more in-depth

description of the most relevant related work has been chosen.

Chapter 4 - Results and Evaluation

In this chapter, results from the research conducted can be found. An evaluation and discussion of the given results and problems of this domain are also presented.

Chapter 5 - Conclusion and Discussion

As the last chapter in part I, we present a conclusion to the work that has been conducted, and what can be further researched.

Part II - Papers

Chapter 6 - Paper 1

This chapter presents the first paper, *On the Challenges of Political News Sentiment Analysis* , in full.

Chapter 7 - Paper 2

This chapter presents the second paper, *Understanding the Political News Domain - Analyzing Negation Count and Co-Occurring Terms in Sentiment Analysis*, in its entirety.

Chapter 2

Theoretical Background

In this chapter, an overview of central terms, theories, and methods used for this thesis, can be found. Section 2.1 describes machine learning theories and methods, and Section 2.2 gives an overview of terms and methods used in the realm of computational linguistics.

2.1 Machine Learning

This section describes central methods and theories, used in the construction of the sentiment engine for this thesis. 2.1.1 accounts for three different machine-learning algorithms, and 2.1.2 describes binary classification which is central for the second paper, found in Chapter 7.

2.1.1 Machine-learning algorithms

Machine learning is the field within Artificial Intelligence that focuses on having computers learn patterns and courses of action over time without being explicitly programmed to do so [5]. By employing classification algorithms on a dataset, given input parameters, it is able to learn from the dataset, and ultimately being able to classify new instances according to the patterns learned. In this thesis, we make use of classification algorithms on an annotated dataset, coupled with computational linguistics described in Section 2.2, in order to create a sentiment analysis engine. Of the algorithms chosen for this task, we have excluded the more computationally demanding ones, such as Support Vector Machines (SVMs) and Annotated Neural Networks [34, 35]. Reasoning for this is because, as described in 1.1, this thesis being the foundation for a module introduced in a system created at NTNU, one that requires as close to real-time computing as possible, and therefore

simpler, more computationally effective algorithms has to be employed. These will now be accounted for in turn.

J48

J48, the Java[1] equivalent to the original C4.5 algorithm developed by Quinlan [36], is a statistical classifier generating decision trees. These are built from a training data, consisting of a set $S = s_1, s_2, ...$ of already classified samples, using information entropy. Each of the elements in S has a vector, $(x_{1,i}, x_{2,i}, ..., x_{p,i})$, where each attribute represents features of the sample and which class it falls in under. Decision trees are generated recursively, where each node are created based on the vector's element that yields the highest normalized information gain. Its formula:

$$IG(A, S) = H(S) - \sum_{t \in T} p(t)H(t) \tag{2.1}$$

where the information gain IG is a measure of difference in entropy before C4.5 splits on an attribute, A. The remaining parts denotes the entropy of set S ($H(S)$), resulting subsets of split on S (T), the proportion of elements in t compared to the samples in S $p(t)$, and the entropy of a subset t ($H(t)$).

Random Forest

Random forest, first introduced by Breiman [37], employs an ensemble of decision trees, constructing a multitude of decision trees during training of the classifier, and outputs the average of all classification trees. It corrects for overfitting, as noted by Hastie et al. [38], and is given by:

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(x^{'}) \tag{2.2}$$

where B is the total number of trees used, and $\hat{f}_b(x^{'})$ denotes decision tree b's classification of instance $x^{'}$.

Naïve Bayes

Lastly, Naïve Bayes, which is the classifier reaching highest precision scores in the Norwegian political news domain for this thesis, is a simple probabilistic model that takes on the assumption of all features, $\langle a_1, a_2, ..., a_n \rangle$, being conditionally independent given a class label [39]. By employing frequencies of features, it

[1] Object-Oriented programming language used in creation of sentiment engine for this thesis.

is able to classify documents (in our case, paragraphs) into its respective class. Formally, its classifier can be given by:

$$v_{NB} = arg \max_{v_j \in V} P(v_j) \prod_i P(a_i|v_j) \tag{2.3}$$

where v_{NB} denotes the target output by the classifier. First, estimation of the probabilities, based on frequencies in training data, are performed. This hypothesis is then applied to new instances being classified. Even if it being a simple method, with enough pre-processing, it can be a viable contender to the more advanced methods such as SVMs [40].

2.1.2 Two-step Binary Classification

Within the approaches of supervised and unsupervised learning, classification tasks is a common technology employed. As outlined in the previous Section, 2.1.1, those algorithms were used for that exact purpose. Commonly, classification between two different, opposing, polarities, or on the continuum between these, takes place, according to Pang and Lee [10]. Given the sentence:

> 500 barn ble brutalt drept i bombinger i Baghdad.
>
> *500 children were brutally murdered in bombings in Baghdad*

which is negative. If following definitions from Balahur et al. [41], it is a negative sentence, but a factual one, not one bearing sentiment. Therefore it should be classified as objective. Occurrences like this, or similar sentences, in the news domain is common. Sentences that seem to be negative, but have an objective, factual focus. To better gain results we employ a two-step binary classification task consisting of the two steps: *subjectivity detection* and *polarity classification*.

To understand if a paragraph is opinionated or not, subjectivity detection is implemented. This to filter out the paragraphs that have an objective focus, like the one in the example above. Mihalcea et al. [42] argues that improvements on this classification task, will ultimately have positive effects on polarity classification. By filtering out the objective paragraphs in the first step, it gives a better focus on distinguishing between the polarities of a paragraph in a second step.

If labelling a paragraph as either overall positive or negative, we refer to this task as polarity classification. Even if not necessarily opinionated, input to such a task usually is. By having a paragraph piped from the first step, and only having to distinguish between two opposite polarities, features can be included that only focus on this task. Features like this will be further explained in Section 2.2.

2.1.3 Classification Evaluation

Once a classifier has been trained, evaluation is needed. This is done to both determine its error in classification, and as well compare it to other classifiers working on the same dataset. Following are the metrics focused on when evaluating said classifiers, as well as one method making use of these metrics.

Classification metrics

In terms of how well a sentiment analysis system is classifying its data, metrics are introduced. Of these, we will focus on *accuracy*, *precision*, and *recall*, all of which are used in this thesis to evaluate the aforementioned classifiers. To visualize results from classifiers, we employ a *confusion matrix*, and will therefore also explain this in short.

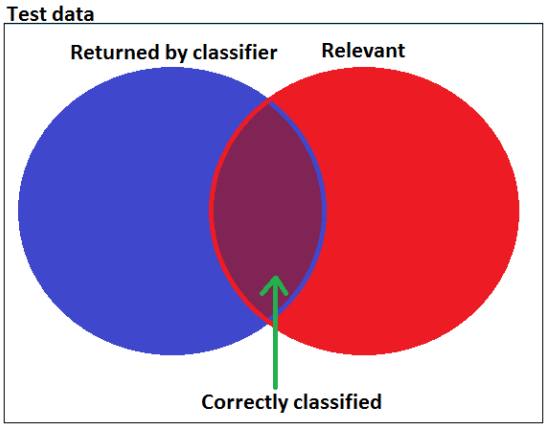


Figure 2.1.1: Venn diagram for binary paragraph classification.

A simple illustration (Figure 2.1.1) denotes the subset of paragraphs returned by the classifier, the subset of relevant paragraphs, and the subset of correctly classified paragraphs. Using these subsets are the equations 2.5, and 2.6. Equation 2.4 also make use of the total number of paragraphs in the test data.

$$accuracy = \frac{\#\text{ correctly classified}}{\#\text{ total test data points}} \tag{2.4}$$

$$precision = \frac{\#\text{ correctly classified}}{\#\text{ total items returned by classifier}} \tag{2.5}$$

$$recall = \frac{\#\text{ correctly classified}}{\#\text{ total relevant items}} \tag{2.6}$$

From these equations, we are mainly focusing on precision to measure the system quality, though recall and accuracy are included to give an overall estimation, and can pinpoint usage areas for a classifier.

Confusion matrix. If only focusing on accuracy as described in equation 2.4, distinguishing of classes are not shown, only the overall accuracy of the classifier. To better understand the data, in terms of how many data points belonging to each class, and where tuning can be performed, class-specific accuracy can be measured from data in Table 2.1.1. In this confusion matrix, the first row represents what the paragraphs was classified as, and last column denotes correctly classified paragraphs (diagonal of highlighted numbers).

Table 2.1.1: Confusion Matrix for hypothetical classifier, EC2

a	*b*	*c*	← Classified as
601	201	195	*a = 1*
76	**1200**	124	*b = 0*
158	700	**403**	*c = -1*

Observed from the matrix, the classifier, *Example Classifier 2.0 (EC2)*, performs better in terms of the neutral and positive class, and worse in regards to the negative class. When considering the false negatives (the portion of incorrectly classified paragraphs to a given class), the value 700 is much higher than any of the others. If tuning is of interest, a focus should be on the negative class, as this has a negative impact on the overall accuracy.

Table 2.1.2: Precision and recall for hypothetical classifier, EC2

Class	Precision	Recall
1	72.0%	60.3%
0	57.1%	85.7%
-1	55.8%	31.9%
Overall	61.6%	59.3%

Table 2.1.2 depicts the different precision and recall scores for each class in classifier, EC2. As noted earlier, the negative class has the worst results. If, for instance a sentiment analysis system is interested in a high precision value for its positive elements, but ignores the lack of returned elements, this classifier is one that could be used. It is often a trade-off between precision and accuracy. If there is a high precision value, subsequently a lower recall value follows, and vice versa.

Cross-validation

In this thesis, evaluation with the use of the *holdout method*, and similarly *random subsampling*, which partitions its data into two mutually exclusive subsets, training and test set, alternatively repeating the process [43], have been omitted due to their chance of overfitting. We have rather focused on *cross-validation*, which makes sure each record is used the same number of times for training, and only once for testing. The goal of a cross-validation process, is to define a dataset which in turn will be used to test the classifier in the training phase. Doing so, several rounds of cross-validation takes place, where in each round, partitioning of the data set into a test and training set is performed. The resulting scores are averaged and gives an overall precision of the classifier. This method works well for small test sets, one that is present in this study.

2.2 Computational Linguistics

Computational linguistics is the field of understanding and modeling natural language to be used by a computer system [44, 45]. To better the understanding of how a language works, in terms of logic, its structure, and how meaning is assigned to text, is what is of interest here. We will now describe in further detail the aspects and methods implemented in the construction of our sentiment analysis engine.

2.2.1 Part-of-speech

Part-of-speech (POS) is a category of lexical items with grammatical properties, as explained by Kroeger [46]. In the English language there are eight word classes, with three additional, less used. The Norwegian has ten [47], which are outlined here:

Adjektiv (Adjective). Describes one or more nouns.

Adverb (Adverb). Describes or modifies verb, adjective or other adverbs.

Determinativ. More thorough description of nouns (no equivalent word class in English).

Interjeksjon (Interjection). Words expressing emotion or sentiment on part of the speaker.

Konjuksjon (Conjunction). Combines two words or phrases.

Preposisjon (Preposition). Describes where a verb or noun is in relation to another verb or noun.

Pronomen (Pronoun). Replacement of noun.

Subjunksjon. Sub-class of conjunctions which initiates phrases (no equivalent word class in English).

Substantiv (Noun). Name of a place, person, or thing.

Verb (Verb). Name of an action.

For implementing these into our sentiment analysis system, we have made use of a POS-tagger, the Oslo-Bergen Tagger (OBT)[2]. This tagger is given input paragraphs and assigns a word class to each word or term in said paragraphs. Of the word classes above, we have put a focus on the four classes, verbs, adverbs, nouns, and adjectives, as these are the ones often bearing a sentiment value [22].

2.2.2 Pre-processing

Pre-processing in linguistics is the process of extracting the conceptual information its text applies to [48]. Work conducted by the same authors suggest improvements when performing pre-processing. Haddi et al. [49] also stresses the importance of pre-processing in order to gain better results. Of the different pre-processing techniques, we will focus on two: lemmatization and tokenization.

Lemmatization

Words such as *jump*, *jumped*, *jumps*, *jumping*, is all a version of the same word - its lemma, *jump*. Lemmatization is similar to stemming[3], though it employs a vocabulary and morphological analysis to return the base or dictionary form of a word [50].

[2]http://www.tekstlab.uio.no/obt-ny/ created by the University of Oslo (UiO) and University of Bergen (UiB). Read on for further documentation.

[3]Stemming is the process of pruning words, though it only removes the end of a word.

Tokenization

Tokenization is the process of splitting a sequence of text into pieces, called *tokens*, followed by possible stripping of e.g. punctuation, leaving us with only the word [50]. These can often be only words, but sometimes it is important to keep the token in full. For example, the word *won't*, if split without accounting for the hyphen, we are left with two elements, *won* and *t*. The first makes sense, in a way, though it does not reflect the meaning of the original word, whereas the second does not make any sense. Tokenization is employed to keep these combinations intact.

2.2.3 Valence shifters

Valence shifters are categories or set of words that change the meaning of a sentence, going from positive to negative, from a weak to a strong sentence, or shifting its sentiment [51]. Work by Kennedy and Inkpen [52] show increase in accuracy in a sentiment analysis system when implementing valence shifters. Here, we will account for five of these categories.

Diminishers. Diminishers are words decreasing the degree of expressed sentiment in a phrase. Diminisher words like *mindre*, *færre*, *såvidt (barely)*, *lite*, will decrease the intensity in a phrase. Take for example, *This movie is barely good*, diminishes the sentiment effect of *good*.

Connectors. Connectors are words that connect two phrases or sentence, to form a longer sentence. These can introduce information, though at the same time often contradict the texts combined. Conjunctions are included as connectors, as well as multi-word constructs. Words and constructs like *og*, *eller*, *men*, *fordi*, *hvis*, *selv om (although)*, *til tross for*, *dessuten* are examples of connnectors. The sentence, *Selv om Kristoffer er en god venn, er han en forferdelig ektemann*, have a positive sentiment in the first phrase, though it is negated by the second.

Intensifiers. Intensifiers, similar, but opposite to diminishers, are the words increasing or intensifying the sentiment in a phrase. Words like *meget*, *kjempe*, *mer*, *veldig (really)*, *bestemt*. Example, *This movie is really good* has a stronger sentiment than *This movie is good*.

Negations. Negations, or negatives, are the most common shifters. Inclusion of a single word that negate the polarity of words, phrases or sentences. Take the sentence, *Kristoffer er en dum fyr* which has a negative sentiment towards Kristoffer. By including a negation, namely *ikke* in this example, the sentence

reads, *Kristoffer er ikke en dum fyr*, which now no longer has a negative sentiment towards Kristoffer. In the Norwegian language, we have considered the words, *ikke*, *ei*, *nei*, *aldri*, *neppe*, *inga*, *ingen*, *intet* with their equivalent dialect versions. The inclusion of negations is central in our second paper, found in Chapter 7.

Verbs. Verbs can also act as valence shifters as these are the words believed to have the strongest weight in terms of sentiment. These can act as diminishers, intensifiers, like in *Snille Kristoffer er dum*, where *dum (stupid)* has a stronger sentiment impact on the phrase.

2.2.4 Co-Occurring Terms

Co-Occurring Terms (COTs) are terms that co-occur in the same context, which also bear importance to each other, as noted by both Pang and Lee [10] as well as Matsuo and Ishzuka [53]. From Matsuo and Ishizuka, creating a sentiment lexicon based on COTs is proposed as a good substitution for a corpus, as further explained in Section 2.2.5. Subsequently, we adhere to the following three definitions in creation of COTs, as well as a special case in the Norwegian language.

Co-Occuring Terms (COTs). *COTs are terms that can occur anywhere in a span of text, though without being separated by sentence-ending punctuation such as periods, question marks, or exclamation marks.*

Take the sentence, *Erna Solberg åpner talen. Er dette noe som vil bli tatt seriøst, eller vil hun bli latterliggjort? Finn ut!*. Following the above definition, COTs like *[Solberg, talen]*, *[tatt, eller]*, *[er, latterliggjort]*, are all possible candidates. In order to reduce the number of total candidates, additional constrains are implemented, namely arity and radius.

Arity. *Arity sets the number of terms a COT can consist of.*

Given an arity of 4, a COT can be composed of the four terms, *det, er, flott, alle*. A larger arity can have a COT make more sense on its own, though a smaller arity will result in creation of more candidate COTs. In this study we have set a default arity of 2.

Radius. *A radius sets the maximum distance between the occurrence of two terms.*

For example, given the sentence, *Bondevik sitter på stortinget, men har lite konstruktive saker å ta opp*, and a radius of 11, any of the words can be combined to make a COT, like for instance *[Bondevik, opp]*. Given a radius of only 2, there are only two possible COTs for *Bondevik*, namely *[Bondevik, sitter]* and *[Bondevik, på]*. In this study we have set the radius to be either of the following, {2,4,6,8}.

'Preposisjonsuttrykk'. COTs are often comprised of two single words, though there are special cases, and especially in the Norwegian language, where the terms in a COT can consist of two or more words. Such a case is with the occurrence of a *"preposisjonsuttrykk"*, that servers as an adverbial or adjectival clause in a sentence [54]. Terms like *[skal, i stedet] (shall, in place)*, where the second term, *i stedet*, does not make the same lexical sense if divided into two different terms. We therefore extend the word *term* in *Co-Occurring Term* to also include *"preposisjonsuttrykk"*.

2.2.5 Lexicon

Lexicons are the vocabulary of a language, an individual speaker or group of speakers - the total stock of morphemes in a language [55]. In sentiment analysis, lexicons can help algorithms better understand the context and meaning of the contents it is applied to, and is also crucial for most of them, including the ones outlined in Section 2.1.1 [17]. Lexicon used in this thesis make use of COTs, and employ the ranking functions and algorithms that will be accounted for here and in Section 2.2.5 respectively.

Ranking functions

In its simplicity, ranking functions are used for ranking the features (in our case: COTs) presented, and give these as output. By making use of top-ranked COTs, a decrease of effective vocabulary size is evident, leading to increased computational efficiency, less manual labor in annotating, and at the same time, excluding COTs that does not bear importance to the overall data set [50]. Here, we will account for five different ranking and bias functions, employed in the acquisition of lexicon used in the sentiment analysis system created for this thesis.

Term frequency. The simplest form for ranking, is by use of the term frequency method. Given a class of text, a weight is assigned to each COT present in that class. On counting the number of occurrences for a COT, this weight is increased, using an integer. What can be changed is the definition of the class, where it can be a collection of documents, or in most cases, an instance of one single document.

Term frequency is denoted as,

$$tf_a \tag{2.7}$$

where a is the COT being weighted. Term frequency's simplicity is one of its strengths, though without pre-processing of which COTs to look for, frequent COTs without specific information can be returned [50].

Inverse document frequency. Following the scenario of not pre-processing which COTs to look for, we are left with one big problem: all COTs are considered equally important. Take the simple sentence,

> Det er ikke det som betyr noe, det er tvert imot det motsatte.

where *[det, er]* (count: 2) would give a higher term frequency than *[betyr, noe]* (count: 1) or *[det, motsatte]* (count: 1), which is more indicative of the contents of the sentence. Similarly, in the political news domain, having the term *politikk (politics)* occur in large quantities of the collection, would mitigate its importance. Inverse document frequency accounts for this. By looking at the number of instances in a collection that contains a COT, a boost is given to these specific instances [50]. Inverse document frequency is denoted as,

$$idf_a = \frac{N}{df_a} \tag{2.8}$$

where a is the COT being weighted, and N is the number of instances in the collection. Following this definition, COTs that occur in a lower number of instances over the whole collection, will yield a higher ranking.

TF-IDF. Combining both ranking methods described above, TF-IDF yields a score proportionally increased by the number of times a COT occurs in an instance, though is offset by the frequency in the collection as a whole. This combined ranking method will therefore rank instances with a unique COT, though occurring more often in said instance, higher than if only focusing on its inverse document frequency. Given the notation,

$$tf - idf_a = tf_a \cdot lg(idf_a) \tag{2.9}$$

where tf and idf refers to the term frequency and inverse document frequency, respectively, as previously described.

Mutual information. Mutual information is a measure of the amount of information one obtained COT, have on the classification of another instance. The mutual information of COT a and instance c is calculated given [50]:

$$MI(e_a; e_c) = \sum_{e_a \in 1,0} \sum_{e_c \in 1,0} P(e_a, e_c) log_2 \frac{P(e_a, e_c)}{P(e_a)P(e_c)} \tag{2.10}$$

where variable e_a denotes if the instance contains the COT or not, depending on the numerical value of 0 or 1. Similarly, the variable e_c, whether assigned the value of 0 or 1, denotes the presence of the instance in class c. Once computation of all COTs a over instances c has been done, one can rank all COTs accordingly, and select the ones with the highest rank, subsequently ending up with a set of the top ranked features.

Chi-squared. Chi-squared is a bias method used in statistics to test the independence of two variables, A and B. They are defined to be independent if $P(AB) = P(A)P(B)$, or equivalently, $P(A|B) = P(A)$ and $P(B|A) = P(B)$. In COT ranking, the two variables are occurrence of the COT a and occurrence of the class c. Given by [50], we can rank COTs according to:

$$\chi^2(a, c) = \sum_{e_a \in 0,1} \sum_{e_c \in 0,1} \frac{(N_{e_a e_c} - E_{e_a e_c})^2}{E_{e_a e_c}} \tag{2.11}$$

where e_a and e_c are defined as in 2.2.5, N the observed frequency and E the expected frequency, assuming both COT a and class c to be independent. This will result in a measure of how much the counts of A and E deviate from each other. Given a high value of χ^2, indicates that the hypothesis of independence is incorrect, with a fairly high percentage of confidence.

Co-Occurring Term Algorithms

The overall method in the acquisition of sentiment lexicon used, depends on the generation and ranking of COTs. The two most important algorithms for this purpose are shown in Figure 2.2.1 and 2.2.2 and are explained below. Even though we are doing our analysis on the paragraph level, the lexicon is created on the article level. As such, when we talk about a *document* in this section, we are referring to an article, **not** a paragraph. The reason for this is that a paragraph is often very short in our domain and thus there is a very low chance of a COT occurring more than once in a paragraph.

Generation of COTs. To generate COT candidates for our lexicon we implement the GENERATE-COTS(D, r) algorithm shown in Figure 2.2.1. This takes two parameters: D, the data set of manually annotated paragraphs, and the radius r. The output is the set of all candidate COTs, denoted as F_{cot}. As the included algorithm shows, it will loop through the complete set of paragraphs and extract the COTs inside the given radius and matching word classes. The permitted word classes are nouns, adjectives, verbs and adverbs.

We have not shown all of the sub-functions in the figure. The function CLEAN(d) will clean a paragraph by tokenizing it and removing all symbols except those that are sentence-ending (., !, ?), which is needed by the algorithm to understand when a sentence has ended. This is because the two words in a COT need to be in the same sentence. POS-tagging with the Norwegian Oslo-Bergen-Tagger is also done here. EXTRACT-WORD-AT(d, i) simply fetches the word at position i in paragraph d. The PERMITTED-POS function checks wether the current word is in one of the permitted word classes mentioned above.

Ranking of COTs. Figure 2.2.2 illustrates the second algorithm we implement, GENERATE-COTS-RANKING(D, r, σ, f), taking four parameters. Again, D, is the set of annotated paragraphs, and r is the radius of the COTs. σ specifies the size of lexicon, and f is the ranking function used to rank our candidate COTs. The output will be a set of σ ranked COTs, F^*_{cot}, which we then annotate to complete our lexicon. Starting with an empty set, we first extract the COTs, which are candidates for annotation. This is done in the GENERATE-COTS(d, r) function, detailed in Figure 2.2.1. Each and every COT is then ranked with the specified ranking function − explained in Section 2.2.5 − and added to a temporary set of ranked COTs, $F_{(w \oplus v)}$. Notice that only COTs that are present in more than 1 document are added as there is very limited benefit of including COTs that occur in only 1 document.

The TERM-FREQ(D, $w \oplus v$) function computes the number of times the COT comprised of words w and v occurs in the whole data set D. The DOC-FREQ functions counts the number of documents a COT or a single word occurs in. This means that multiple occurrences in the same document are not counted here. We have omitted the computation of term frequencies of single words as these are not needed in any of the chosen ranking functions. COMP-STAT takes the ranking function f and all the computed frequencies as input and returns the ranking of the current COT $w \oplus v$ based on f. SORT-DESC simply sorts the ranked COTs in descending order so that only the COTs with the highest rankings are included.

2.3 Statistical Methods

Statistical methods is the process of collecting variable numerical data relevant to what is asked for, summarizing using appropriate statisticical calculations and display information gathered, and analyze the observed data according to theoretical models [56]. In our first paper (Chapter 6), when manually annotating the same dataset, the reliability between the different annotators has to be analyzed. This is done by measuring the level of agreement, the inter-rater reliability, between the annotators. We focus on one method for this process, namely the joint-probability of agreement.

2.3.1 Joint-probability of agreement

The simplest form for inter-rater reliability is the joint-probability of agreement. Given the dataset to be annotated, for each paragraph annotated as one of the classes (positive, negative, or neutral), a numerical value is added to that class. The total numerical value is then divided by the number of annotations. In simple English, if all annotators has annotated a paragraph as positive, there is a 100% agreement for that paragraph. This is then aggregated over the whole annotated dataset, resulting in a total percentage score. There are obvious downsides to this method, as agreement by chance is likely, suggested by Uebersax [57]. The annotators do, however, follow the same guidelines, an agreed-upon sample of the ones proposed by Balahur and Steinberger [41], and therefore see the joint-probability of agreement as a valid measurement method.

Input: D: paragraph data set, r: radius of COTs
Output: F_{cot}: complete set of COTs

```
1: F_cot ← {∅}
2: for d ∈ D do
3:    d ← CLEAN(d)
4:    for i = 1 to LENGTH(d) do
5:       j ← 1
6:       w ← EXTRACT-WORD-AT(d, i)
7:       if not PERMITTED-POS(w) then
8:          continue
9:       end if
10:      while j ≤ r and v ∉ {., !, ?} do
11:         v ← EXTRACT-WORD-AT(d, i + j)
12:         j ← j + 1
13:         if not PERMITTED-POS(v) then
14:            continue
15:         end if
16:         if (w ⊕ v) ∉ F_cot then
17:            F_cot ← F_cot ∪ (w ⊕ v)
18:         end if
19:      end while
20:   end for
21: end for
22: return F_cot
```

Figure 2.2.1: GENERATE-COTS(D, r): Algorithm for the generation of COTs from a data set D of paragraphs with a radius of r

Input: D: paragraph data set, r: radius, σ: size of lexicon, f: ranking function
Output: F^*_{cot}: ranked set of COTs

$F^*_{cot} \leftarrow \{\emptyset\}$
$F_{cot} \leftarrow$ GENERATE-COTS(d, r)
for $(w \oplus v) \in F_{cot}$ **do**
  $df_w \leftarrow$ DOC-FREQ(D, w)
  $df_v \leftarrow$ DOC-FREQ(D, v)
  $tf_{(w \oplus v)} \leftarrow$ TERM-FREQ$(D, w \oplus v)$
  $df_{(w \oplus v)} \leftarrow$ DOC-FREQ$(D, w \oplus v)$
  if $df_{(w \oplus v)} \leq 1$ **then**
    continue
  end if
  $F_{(w \oplus v)} \leftarrow$ COMP-STAT$(f, df_w, df_v, tf_{(w \oplus v)}, df_{(w \oplus v)})$
end for
$F^*_{cot} \leftarrow$ SORT-DESC(F_{cot}, F)
$F^*_{cot} \leftarrow F^*_{cot}(1 : \sigma)$
return F^*_{cot}

Figure 2.2.2: GENERATE-COTS-RANKING(D, r, σ, f): Algorithm for the ranking of a list of σ COTs, from Figure 2.2.1, based on the data set of paragraphs D, radius r and ranking function f

Chapter 3

Related Work

In this chapter, we present related work to our thesis. We will, for the sake of not repeating detailed sections in each of our papers, found in Chapter 6 and 7 respectively, only focus on some of the related work, the main categories that relates to our paper as a whole.

In terms of sentiment analysis, a growing field of research, there are a lot of related work to our thesis. If pinpointing it down to the niche of political news, and especially news articles, there are considerably less. We therefore recognize the two main categories that is closely related to our work, namely 1) the political language and its rhetoric, and 2) sentiment analysis of political text, difficulties, how news articles compare to microblogging posts, and how already-existing methods can be employed. We will now account for these two categories in turn.

3.1 Political Language

Political language is one that often does not convey its sentiment in a simple way. Rather, as suggested by Rozina and Karapetjana [58], being linguistic manipulative. Wodak [59] defends this statement, noting the political language to have persuasive grammar, employing use of phonetics signaling words that in truth have different meaning, and changing sentiment with the use of contextualisation. Edelman [60] signals evocative rhetoric being employed in the same language. Bell [16] delves further into the language used by news media in general. The political language is believed to be difficult to understand for the everyday person, according to Grottum et al. [61], and is therefore seen as challenging for a sentiment analysis system to account for. The language used by journalists, also in political news, is an objective language, refraining from using subjective vocabulary, instead shifting focus to quotes et cetera that portray sentiment [21].

3.2 Sentiment Analysis of Political Text

Up until this point, most research in terms of political text has been done on the microblogging platform Twitter[1], and is usually concerned about prediciting electoral results [62, 63, 64, 65, 66, 67]. Research by [68] highlights the limits for using Twitter as a predictor for this. In terms of sentiment analysis in political news articles, work as that of Mourad and Darwish [69] show promising results. Similarly, Kaya et al. [70] employ machine learning and sentiment lexicon in the Turkish political news domain.

Sentiment analysis is commonly formulated as a ternary classification problem. Text is grouped into either of the classes, positive, negative, or neutral. In such classification tasks, sentiment lexica are usually employed [17]. Research by Njølstad et al. [33] incorporate these methods, the same ones as in this thesis, for classifying news articles in the financial news domain.

Introduced in later years are the methods of subjectivity and polarity classification. By employing steps where in the first, text with subjectivity, is detected, and in a second step, the polarity of the text is determined. Research by Wilson et al. [32] employs this method of two steps and achieved state-of-the-art results. Similarly does Mourad and Darwish [69].

[1] https://about.twitter.com/

Chapter 4

Results and Evaluation

In this chapter we present the results obtained in our research. These results are mostly duplicated from our papers (Chapter 6 and Chapter 7), though we feel it is necessary to include it all here to show how the papers are connected, and to be able to conduct a detailed discussion on the results in our domain. As our domain has shown to be more complex than those we compare with, this discussion is significant to our research contributions, and hence is just as important as the isolated results. We start out by describing our dataset with some statistics in Section 4.1 before going into presenting the precision scores from both studies in Section 4.2. These score are compared to the results from the financial news domain illustrated in [33], and other studies mentioned in "Related Work" (Chapter 3), which makes up what is called the state-of-the-art classification in sentiment analysis. Subsequently we delve into our theories about the characteristics of the political news domain contributing to the results we achieve with the used methods in Section 4.3.

4.1 Dataset

To better understand our results and subsequent discussions about them we include a description of the dataset we have used in training and testing of the machine learning classifiers. As mentioned in both our papers, the dataset is composed of news paragraphs from NRK and VG. In total, the annotators annotated 3961 paragraphs, which resulted in a joint-probability of agreement of 76% (3016 paragraphs were then used for training and testing). The average length per paragraph is 27.3 words. Table 4.1.1 illustrate that the number of neutral paragraphs are more than double that of positive, and ∼50% more than the number of negative. This is clearly unbalanced in a ternary world, but it also suggest that our

Table 4.1.1: Annotated paragraphs by class

Positive	698
Neutral	1426
Negative	892

Table 4.1.2: Average COTs per paragraph

Positive	0.43
Neutral	4.65
Negative	0.24
Total	5.3

dataset is close to balanced when it comes to number of sentiment-bearing paragraphs and objective paragraphs, which are the two classes in the subjectivity classification step discussed in Section 4.2.2.

The total number of annotated COTs depends on which lexicon parameters are used. Table 4.1.2 details how many COTs from each class there are per paragraph in our dataset when the lexicon employed is of size 4000, the radius is 4, and the ranking function is TF-IDF. There is a significant difference in number of neutral COTs per paragraph compared to the other two. While there is less than one COT per paragraph for the subjective classes, the neutral class has more than four per paragraph. This is one of the reasons why the political news domain is so challenging, which is something the results and discussion in the rest of this chapter will explain in more detail.

4.2 Sentiment Engine Construction

The promising results shown in research conducted by Njølstad et al. [33] in the domain of Norwegian financial news justifies the replication of their methods. In addition, this would solve our problem with the limited lexical and linguistic resources in the Norwegian language, with a reasonable amount of manual labor to annotate COTs in the sentiment lexicon. Paper I details how we built our sentiment lexicon, which we then used as a tool in feature extraction, and discusses possible causal factors behind our dissatisfying precision scores. Paper II is thus a consequence of these scores as we investigate ways to improve the results and understand the domain better without discarding our sentiment lexicon.

Table 4.2.1: Classification precision of a ternary system with input parameters $(r \times \sigma \times f \times c)$.

		σ = 1000				σ = 2000				σ = 4000			
MLC	Ranking	COT radius =				COT radius =				COT radius =			
	Function	2	4	6	8	2	4	6	8	2	4	6	8
NB	*tf*	54.9	55.1	54.9	54.0	56.8	57.3	56.3	55.1	53.6	57.4	57.9	56.6
	idf	52.4	52.5	52.0	51.4	50.8	53.6	52.6	52.9	53.6	51.6	51.3	51.7
	tfidf	55.3	55.2	54.6	53.9	56.8	57.5	56.8	54.9	53.7	57.2	57.9	56.6
	mi	51.9	51.3	50.8	50.9	52.3	52.5	52.1	52.8	53.6	53.5	53.8	52.9
	χ^2	53.3	52.1	52.2	50.8	53.5	52.8	53.1	52.4	53.6	54.3	52.6	52.8
RF	*tf*	51.3	51.1	51.3	48.6	52.6	52.3	52.0	50.7	50.2	54.5	53.1	53.2
	idf	49.9	48.1	48.2	49.1	47.8	49.2	47.9	48.2	50.3	47.5	47.2	48.2
	tfidf	52.2	51.5	51.3	48.5	52.8	52.5	52.4	50.7	50.3	54.2	53.3	52.5
	mi	48.8	48.9	47.2	47.3	47.8	50.3	48.8	47.7	50.3	49.6	49.1	49.0
	χ^2	47.2	49.4	48.6	48.0	49.1	47.8	47.1	48.8	50.3	50.5	46.7	47.1
J48	*tf*	56.0	55.7	55.9	54.3	57.4	58.0	56.3	56.2	54.8	59.0	59.5	58.8
	idf	51.6	51.5	51.1	50.6	50.3	52.5	51.4	51.8	54.8	50.0	51.8	52.0
	tfidf	55.6	56.0	55.4	54.5	57.5	57.8	56.6	56.1	54.7	**59.7**	59.5	57.7
	mi	51.2	51.1	49.3	51.3	53.7	52.7	53.0	52.5	54.8	53.6	54.7	53.1
	χ^2	53.4	51.8	51.0	49.8	53.3	53.3	52.9	51.9	54.7	53.2	52.5	54.3

4.2.1 Ternary Classification

To build our sentiment lexicon we annotated ∼4000 paragraphs from political news articles gathered from NRK and VG, which gave us ∼10,000 unique candidate COTs. These were then ranked according to the ranking functions described in Section 2.2.5. Combining the ranking function (f) with the COTs parameter radius (r) and a lexicon size (σ), all the resulting lexica were tested with the three machine learning classifiers (c) we detailed in Section 2.1.1. The results of this experimentation is shown in Table 4.2.1 with all the possible combinations from $(f \times r \times \sigma \times c)$. For more details about this work we refer to Paper I (Chapter 6).

The classification precision results illustrated in Table 4.2.1 shows us that we were not successful in achieving state-of-the-art precision scores, which are upwards of ∼70% [21, 32], for any parameter combination. More importantly, we are not in a satisfying range of Njølstad et al. [33] either, where the most comparable parameter combinations yielded a high of ∼65%. In bold is the best variable combination in our system, with a precision of 59.7%. Considering the overall picture from our results, however, we can observe some of the same trends as Njølstad et al. reported. Specifically, a smaller radius means a stronger relationship between the terms in the COT, as the precision is in general declining as the radius gets bigger for each lexicon size. Exceptions could be due to the coupling of radius and lexicon size when using an absolute size as we talk more about in Paper I. Further, as the lexicon size increases, so does the precision. However, we do not see any declining increase in precision as in Njølstad et al. We theorize that this is due

to the low precision scores, which can contribute to a more random distribution when many of the combinations show results close to a guessing game. In addition to this, we also see that the machine learning classifier, J48, yields the highest precision, and that the ranking term *frequency-inverse document frequency (tfidf)* is the best. In essence we believe these results to stem from the fact that our domain is more complex, and thus cannot rely on a sentiment lexicon based on COTs alone to achieve state-of-the-art performance results. We will come back to this in Section 4.3.

4.2.2 Two-step Binary Classification

From the evaluation and discussion of the results in Paper I, we deemed it necessary to study our domain closer to uncover which features contributes the most to precision during the different phases of sentiment analysis. This meant we had to construct our sentiment engine to do binary classification in two separate steps. Such a system is explained in Section 2.1.2. Implementing this technique allowed us to analyze the features we chose to investigate in Paper II in terms of their effect on subjectivity and polarization respectively, which makes it easier to see where and how to improve the task of classifying sentiment in such a complex domain as political news.

Table 4.2.2: Subjectivity classification: Precision results.

	PosCots	NeutCots	NegCots	Negations	Precision
NB	✓	✓	✓		**67.3**
	✓	✓	✓	✓	66.4
				✓	59.9
RF	✓	✓	✓		62.8
	✓	✓	✓	✓	64.1
				✓	59.8
J48	✓	✓	✓		67.2
	✓	✓	✓	✓	67.0
				✓	61.8

For the experimentation done in Paper II, we focused on the features using our sentiment lexicon, and the use of negations. Observations during closer examination of the dataset led to an analysis conducted of negation count per paragraph in each annotated class, was the background for using negations as a feature in this study. From this we theorized that the negation count in a paragraph could have a positive impact on determining subjectivity and polarization. An example of a negatively annotated paragraph we observed is shown below, where the negation word "ikke" (*"not"*) occurs five times:

> At parlamentarikere som Bøhler er betenkt over at mange, **ikke** minst unge, **ikke** bryr seg om "politikk og sånn", og **ikke** følger med på det

> Bøhler og hans kolleger holder på med, overrasker **ikke**. Det må være temmelig irriterende å vie sitt liv til det parlamentariske demokratiet og så oppleve at folk gir blaffen, kanskje **ikke** engang tror på det.

As Table 4.2.2 illustrates, this was not the case for subjectivity. The detection of subjectivity proved higher precision results without the negation count, only relying on the COTs counts. Another interesting observation is that with this binary classification step, the best performing machine learning classifier proved to be Naïve Bayes, as opposed to J48 with ternary classification in Paper I.

A few more combinations were experimented with during polarity classification. Since we are determining paragraphs as either positive or negative, it seems obvious that we should also use a combination of features excluding the count of neutral COTs. Examining Table 4.2.3, we can clearly see that with every classifier the best combination of features excludes the neutral COTs. The use of negations do not seem to have much impact in this classification step either, and Naïve Bayes is still the best performing machine learning classifier.

The exclusion of neutral COTs is significant since the only way to do this is to use a two-step binary classification process instead of a ternary classifier, where we can separate the classification of subjectivity and polarization, which was one of the reasons for doing this in the first place. As for the precision scores for both steps, subjectivity and polarization, 67.3% and 72.9% respectively, they are both above the performance of the ternary classifier. However, as a binary classification task is simpler, and our results are still a little lower than the state-of-the-art precision for binary classification, we have also put focus on the discussion part in this study. A more detailed description of the work, and the explanation of the negation analysis, can be found in Paper II (Chapter 7).

4.3 Political News Characteristics

The discussion and understanding of the political news domain in terms of the results from our studies is a significant part of our contribution to the field of sentiment analysis, and we have therefore included the discoveries we believe to be most essential to our thesis here. By observing our results in more detail we found factors which defends our statement on how complex this domain is, with the most influential one being the nature of the political journalists and how they convey sentiment in news articles. A confusion matrix in Paper I, the evidence that sparked this discussion, told us that a convincing amount of false negatives for the two subjective classes, positive and negative, were classified as neutral (as opposed to the opposite polarity). In other words, our sentiment engine was not able to understand the existence of sentiment in a majority of paragraphs. When

Table 4.2.3: Polarity classification: Precision results.

	PosCots	NeutCots	NegCots	Negations	Precision
NB	✓	✓	✓		72.7
	✓	✓	✓	✓	68.1
	✓		✓		**72.9**
	✓		✓	✓	69.0
				✓	57.4
RF	✓	✓	✓		64.5
	✓	✓	✓	✓	64.4
	✓		✓		66.2
	✓		✓	✓	65.0
				✓	55.6
J48	✓	✓	✓		72.0
	✓	✓	✓	✓	71.9
	✓		✓		72.4
	✓		✓	✓	72.0
				✓	—

we compared our annotated data set with the data set classified by the sentiment engine, we quickly observed several examples to what we firmly believe to be the two most prominent underlying factors that makes a sentiment lexicon built on COTs seem almost useless. A couple of these paragraphs are included in this discussion to highlight how difficult it is to classify paragraphs in our domain with the use of a COTs based sentiment lexicon.

1. Examples of misleadingly objective paragraphs:

   (a) Venstre-representanten mener bestemt at filmen "Trygghet i hverdagen", som kostet 50.000 kroner å lage, er en valgkampvideo for Frp, ikke en informasjonsvideo for Justisdepartementet.
   The representative for the Liberal party firmly believe that the film "Security in everyday life", which cost 50,000 kroner to create, is a campaign video for Frp, not an information video for the Department of Justice.

   (b) – Den nye forskriften betyr at i vannene der friluftsfolket fortsatt kunne få være i fred, i en liten bortgjemt vik, der er det nå åpent for skuterfolket å bruke turfolket som rundingsbøyer, sier Bjørn Hansen, på vegne av Naturvernforbundet i Finnmark.
   The new regulation means that in the waters where the outdoor people could have piece and quiet, in a small hidden inlet, it is now open for jetskis to use the outdoor people as human buoys, says Bjørn Hansen, on behalf of the Nature Conservatory of Finnmark.

2. Examples of opposing sentiment:

   (a) Naturvernforbundet frykter nå økt trafikk og flere ulykker. Men Frank Ingilæ, leder av Østfinnmark Regionråd, har kjempet for lovendringen, og ser saken på en annen måte.
   The Association for Environmental Conservation now fear increased traffic and more accidents. However, Frank Ingilæ, the leader of Østfinnmark Regional Council, have fought for a change in legislation, and sees the case in a different way.

   (b) Riksrevisjonen mener noen av tiltakene som er varslet er positive, men de er ikke overbevist om at tiltakene holder for å skape et tryggere samfunn: "Etter Riksrevisjonens oppfatning er det imidlertid for tidlig å fastslå om tiltakene vil få den tilsiktede effekten på samfunnssikkerhetsarbeidet."
   The Auditor General believes some of the announced initiatives are positive, but they are not convinced that the initiatives are enough to create a safer community: "The Auditor General's perception is that it is too early to determine if the initiatives will have the intended effect on the community safety work."

The first quote, 1(a), in the list above is also found in Paper I. This example was annotated as negative (-1) by both annotators on these grounds:

- Spending government money to create a campaign video for a political party and then claiming it to be an information video for the Department of Justice is unacceptable
- It does not matter which party the reader support because this is the belief of another politician, which is negative
- This belief is described with an intensifier, "bestemt" (*"firmly"*), which makes the belief more negative
- Including the actual cost of the video, which is tax payer money, in a bi-clause, sets a focus on the fact that the target of the paragraph did something they should not have done

Clearly, none of these reasons has anything to do with negative words or COTs. In fact, there is no negative words or COTs in the paragraph at all. The name of the video is the only place where we can find a subjective COT. However, the COT ["trygget", "hverdagen"] (*"safety"*, *"hverdagen"*) seems to be positive, and is correctly left out of the lexicon due to its low document frequency. Hence, it makes

sense that our sentiment engine classified this paragraph as neutral (0), when the main deciding factor is the number of positive, neutral and negative COTs, and the neutral COTs count is the winner. A very similar behavior is found in 1(b), but we will not go into detail here.

Example 2(a) in the list is another example of a case where a COTs-based sentiment lexicon is not sufficient to accurately classify sentiment on paragraphs. Here we have a paragraph with two sentences. The first one has a negative sentiment. Our classifier would most probably also classify it as negative because of a COT involving the term "frykter" (*"fear"*). The latter sentence on the other hand is more challenging. Our classifier tags it as positive because of the COT ["kjempet", "for"] (*"fought"*, *"for"*). As a part of the complete paragraph, it makes sense that this sentence is positive as it is opposing the first, which is negative. The question then remains: what should this paragraph be classified as? We might argue that it is neutral since we have two opposing sentiments in one paragraph. There is also the same amount of negative COTs and positive COTs. The annotators have classified it as negative based on the fact that they felt the first part carried more weight, and as such the paragraph was incorrectly classified in the end. A study of patterns in paragraphs, and which parts carries the most weight was outside the scope of this thesis, but we include it here to illustrate the difficulties of classifying paragraphs in the political news domain as we have observed the use of opposing sentiments in the same paragraph being frequently used. The last example in our list, 2(b), behaves in the same way. However, it will not be explained here.

From our observations we believe that cases like the examples above are occurring in the Norwegian political news domain too often for a COTs-based lexicon to be the only deciding feature in a sentiment engine. More from this discussion can be found in Paper I (Chapter 6).

As we have previously mentioned, we did not totally discard the COTs-based sentiment lexicon based solely on these results and observations, but rather started experimenting with combining it with other features. Even though our two-step binary classification method did not improve with the addition of the negation feature we introduced in Section 4.2.2, it did shed some light on what the sentiment lexicon does well. We turn our focus to the individual class precision and recall results from the feature combination that achieved the best overall precision.

Table 4.3.1: Step I: Precision and recall for subjectivity classification, NB, tf-idf, with use of COTs.

Class	Precision	Recall
0	57.4%	84.8%
± 1	**76.1%**	43.5%
Overall	67.3%	63.0%

Subjectivity detection is a complicated task, and as we can see in Table 4.3.1, the precision for the neutral class is not very reassuring. Its recall on the other hand is promising. Most interesting here is the precision score highlighted in bold. 76.1% of the paragraphs classified as subjective (positive or negative) is correctly classified as such. This means that in a live system, where these paragraphs would subsequently be the input of the polarity classification step, 76.1% of them would actually belong in this step, and thus the results would improve. In addition, these results lets us understand which parts of the classification process are in need of improvement. Increasing the recall rate of the subjective class would further improve results in the next step, and should be a focus in the future.

Table 4.3.2: Step II: Precision and recall for polarity classification, NB, tf-idf, with use of COTs, except neutral.

Class	Precision	Recall
1	**76.1%**	53.3%
- 1	70.4%	**86.9%**
Overall	72.9%	72.1%

Table 4.3.2 illustrates the individual precision and recall rates in the polarization step. Again, the highlighted numbers show promise. This step does a good job of correctly classifying negative paragraphs, and has state-of-the-art precision results for the positive class. Recall that this is done by excluding the use of neutral COTs, which means that the deciding features are the number of positive and negative COTs. Further, we can theorize that the difference in negative and positive COTs is greater for negative paragraphs. It also indicates that to improve this step, it would be smart to investigate which factors that actually makes a paragraph positive, which negative paragraphs do not have. In other words, improving the recall rate of the positive class. Again, we refer the reader to Paper II (Chapter 7) for more details of this discussion.

[This page is intentionally left blank.]

Chapter 5

Conclusions

This chapter presents our conclusions on the work done in this thesis. Section 5.1 gives a summary of the contribution of thesis to the field on sentiment analysis in the political news domain, while Section 5.2 looks to the future and discusses interesting venues for further research in this field.

5.1 Summary of Contributions

Ultimately, in this thesis we wanted to work with sentiment analysis and understand the political news domain − a challenging domain for sentiment classification compared to other domains. Applying standard methods from well understood domains, such as the financial news domain, and experimenting with new domain-specific approaches and techniques, we have gained knowledge of the domain's characteristics and the effects of these on the sentiment classification task. The motivation for us to choose this domain for our research is partly due to the lack of earlier work with political news, and the more extensive research that has been done on other platforms like Twitter and Facebook. While these platforms form a great measure of people's opinion of politics, which have been shown to increase in importance during political elections, the news world is often the source behind the scenes of these opinions. As we mentioned in Section 1.1, the mass media is setting the agenda of what everyday people should focus on, and this means that it has significant influence of what people think about different politicians and parties. Hence, the massive amount of information from the news deserves attention. We have talked about the problems we face in this domain, but if we could gain a deeper understanding of the domain this would be the groundwork for successful sentiment analysis applications for political news, and might also uncover new methods which can be replicated in other complex domains. As a product of our research questions, formed and explained in Section 1.3, we have

three main contributions.

In response to our first research question, we have uncovered characteristics of our domain, which we believe to have a huge impact on how sentiment analysis can be done. In Norwegian political news, journalists try not to be too polarized, so they often hide their sentiments instead of explicitly expressing them. Compared to other domains, like product reviews, sentiment words (polarized words) are not used as frequently. Instead, sentiment is conveyed by employing more complex sentence structures and contextual knowledge. This causes the sentiment classification task to become very challenging as we need to find domain-specific methods to handle this complexity. Other characteristics involve the lack of formal organization in the news articles, and a tendency to express different sentiments in a single paragraph (and sometimes even the same sentence), which could also be a way for journalists to seem more objective than they actually are. These findings helped us make sense of our results in Paper I (Chapter 6) and were the motivation for Paper II (Chapter 7), which are the basis for answering the two remaining research questions.

Our first sentiment engine, which was a ternary classifier based on the methods and techniques employed in Njølstad et al. [33], achieved results that defends the existence of the characteristics we found in the political news domain. In reply to the second research question, we applied a proven method from another, better understood domain, to our data set with Norwegian political news paragraphs. Our engine yielded results ∼5-10% lower than those we compare ourselves with. This is a significant decrease in performance, which indicates that just replicating methods and techniques from other domains is too simple and does not work well enough in our domain. We do, however, argue that these methods could still be useful, just not as the sole deciding one. Our two-step binary classifier proved great results in the polarity classification step with only the use of positive and negate COT counts. Here we achieved 72.9% in overall precision before any tuning to the machine learning classifier was done (as it is outside the scope of this thesis), which is close to the state-of-the-art performance in binary classification.

The second paper we have included in our thesis details the work on our two-step binary classifier. This work was an experimentation to find solutions to improve sentiment analysis in the political news domain with domain-specific techniques, which is a response to our third research question. To improve it, we must understand it, and this is partly the reason for the binary classifier. Dividing the process into two steps makes it possible observe the specific impact of each feature. In our work, we experimented with a negation count. This feature did not have any positive effects on the results, only a marginal decrease in precision was found in each step. Although we were unsuccessful in improving our sentiment engine with this new feature, we can clearly say that the difference in average negations

per paragraph between the two classes involved in each step is not big enough to have an impact. This could be something to investigate for other domains though. What actually showed to be the improving factor was the exclusion of neutral COTs in the second step as noted earlier. Hence, we argue that proven methods used in other domains could be altered to fit our domain and perform well. Another way to improve precision results is to look at the detailed results from the subjectivity classification. These results are advocating the improvement of the precision of the neutral class, and the recall of the subjective class. Focusing on finding features that divides these two classes will show a momentous increase in overall performance.

We have tried not to repeat ourselves too much in this summary of contributions, and so we have kept it short and to the point. More details can be found in our papers in Chapter 6 and 7.

5.2 Future Work

As an extension of our work, we will discuss how our research and results we have achieved can be taken further in the future. These are ordered according to our research questions:

- Our first research question deals with the characteristics of our domain and how they impact the sentiment analysis task. The aspects we have found all make it more difficult to achieve state-of-the-art precision results. A natural next step here is to categorize the properties of the domain and look for patterns in the data set in terms of sentiment classes. This would be a tremendous first step in the understanding of the domain, which further leads to performance improvements.

- Answering the second research question we replicated a proven method from the Norwegian financial news domain, which is deemed to be more structured and a simpler domain to do sentiment analysis in as it is better understood. It would be too easy to conclude that using standard methods from other domains is fruitless work, only due to the fact that we achieved such inferior precision results. A more preferred way to address this behavior is to analyze other methods and other domains. We experienced a small, but significant, increase in precision by adapting a technique from another domain. This could mean that understanding the effects of other standard methods and techniques in the political news domain paves the way for new adaptations and more prominent gains in performance.

- Improving the analysis of political news sentiments has to be done with domain-specific methods in some way or another. Even though we advocate the use and understanding of methods from other domains, we also believe it is imperative that these are adapted to the domain in question to optimize performance. Developing techniques to make use of the understanding of the language and its complexity, and the unique characteristics of the domain will definitely increase precision. The next step in this direction should be to experiment with high impact features in an environment where it is simple to pinpoint the area of effect. These features obviously need to be found from analysis and categorization as we talked about as future work for the first research questions above.

Part II

Papers

[This page is intentionally left blank.]

Chapter 6

Paper I

Patrik Fridberg Bakken, Terje André Bratlie, Jon Atle Gulla: *On the Challenges of Political News Sentiment Analysis*, submitted to the 4th International Conference on Statistical Language and Speech Processing (SLSP 2016).

On the Challenges of Political News Sentiment Analysis

Patrik F. Bakken, Terje A. Bratlie, and Jon Atle Gulla

Norwegian University of Science and Technology,
Department of Computer and Information Science Tronheim, Norway
bakken.patrik@gmail.com,terje_ab@hotmail.com,jag@ntnu.no

Abstract. Sentiment analysis is a relatively new field of study that has seen a surge of research in the recent years. This has mainly been conducted in structured domains with explicitly expressed sentiments. Examples being Twitter posts, product reviews, and financial news. In political news, a more complex domain, less research has been done. This justifies further experimentation to better understand how sentiment is conveyed in such a domain. Results show challenging characteristics as sentiment is often expressed in a complex and subtle way. Thus, simpler approaches that show promising results in other domains cannot achieve the same precisions by itself in the political news domain.

1 Introduction

Sentiment analysis, often also referred to as opinion mining, is the process of gathering and identifying opinions and view-points from a span of text. A typical application is then to classify these as either positive or negative, and is something that has become even more popular with the surge of the World Wide Web (Internet) and the uncountable number of texts it makes readily available [13], [19], [20]. Especially as people seek towards the Internet to garner information on what other people think about a certain matter, companies within the field of sentiment analysis has taken huge steps to make use of this information. Product reviews are the most common ones, but even the need for political information is an important factor [20]. This can be seen by the research report posted in [25], where as much as half of all American adults went on-line to get news or information about the 2010 midterm elections, or to get involved in the campaign in one way or another. However, even if users go on-line to gain knowledge about a political aspect, it is far more difficult to extract the correct sentiment from news articles than from product reviews. This is because journalists, in most cases, want to appear objective, and if sentiments occur, it is often in a more complex structure than would be found in a product review. A news article, and especially a political news article, is seldom focused on one object, but instead having broader coverage, usually mentioning more than one politician or party when discussing a matter [1].

There are two textual aspects to consider when reading a political news article: the opinion of the journalist, and quotations from politicians. The former,

the language used by journalists, can often be difficult to understand for the general human. A language maybe too comprehensible, one that requires the electorate to have substantial political knowledge before reading the articles, as noted by Grottum et al. [9]. The latter involves often language not used in everyday speech, a language more influenced by allusion, metonymy, and metaphors - the political rhetoric [23].

Foundational research in the field of sentiment analysis has primarily been completed in the English domain [19], [21], [27]. The Norwegian domain has been the subject of far less research. One possible explanation for this is because Norwegian is only spoken by about 4.7 million people in the world, whereas English is spoken by 335 million [5]. It is, as of now, limited lexical resources for a language the size of Norwegian [22]. One study completed by Njølstad et al. [18], focuses on the creation of a sentiment lexicon from co-occurring terms (COTs) and manual annotation of documents in the Norwegian financial news domain.

A research project at the Norwegian University of Science and Technology (NTNU), Smart Media, have work related to linguistics and news recommendations [10], [26]. Work in the financial domain, and the contributing factor to the Smart Media project, is one that has given motivation for this research.

The focus in this paper is therefore two-fold. We first make use of similar methods as those in [18] to create a domain-specific sentiment lexicon for the Norwegian political news domain. As the political domain is believed to be a more unstructured one [1], we hypothesize that it is also more difficult to gain high precision, mostly down to the subtle way political sympathies are expressed. The second contribution will therefore be to see if similar results in terms of precision can be replicated, and discuss problems in regard to this.

2 Related Work

In terms of political news, most research has been completed on Twitter posts, such as [4], [28]. Sentiment analysis in news articles have been of little focus, and especially in political news. Initial efforts is work completed by Evgenia et al. [6], and Blaz et al. [3].

As previously noted, sentiment analysis identifies the opinion on a span of text. This can either be a full document, or parts of a document, such as a paragraph, sentence, phrase, or word. Most research has been done on the document-level [7], [29], where often an aggregate of sentiment tells us if the entire document is positive or negative. In the political news domain, when taking into account a whole article, there are often several targets (e.g. politicians and political parties). To aggregate all sentiment scores in a full article can therefore prove problematic and not so interesting, especially if there are differences in polarity towards the different targets. On a sentence-level, several targets are not so frequent [7]; but there can also be occurrences of sentences without a target. To ensure that a target and an opinion are present in the span of a text, we have moved up to the paragraph-level, though at the same time treating this as

a small document. Less research has been done on the paragraph-level, though the research conducted in financial blogs [8] suggests that the precision scores may be higher than for sentiment analysis at the document level and thus gives reason for a better precision than when focusing on the article as a whole.

As we treat a paragraph as a sub-document, the same approaches apply to our research. Hereunder, there are two main approaches, both making use of learning algorithms: supervised learning and unsupervised learning. Supervised learning can be as simple as having two classifications: negative and positive. Provided a training data set, and employing a common classification algorithm such as Naïe Bayes (NB), a model is learned. This in turn is used for classifying new documents into one of the classifications described. On the other hand, unsupervised approaches makes use of semantic orientation on phrases in the document, and given the average of these a document is then either classified as positive or negative [7].

Noted by Lu et al. [14], the meaning of a word is highly contex-dependant. It is therefore common, and also seen as imperative, to include the use of a sentiment lexicon when using an unsupervised approach as described above. Creating a lexicon, one out of three methods are commonly used: manual coding of the lexicon, seed word-based approaches extracting words by use of already-made resources, or a large domain-specific corpus [7].

Since both supervised and unsupervised approaches have their limitations [20], research has been done on combining the strengths from both of these, as described in [16]. By first using a lexical tool to label examples (training set), one can train a classifier based on this labelled training set. This has shown to outperform the single-type approach of only using either a supervised or unsupervised one. Completing the label process of a training set does not yield as good results as one would think, and is something that outlines the difficulties in sentiment analysis. From previous studies in different domains, such as [12] and [29], human agreement ranging from 63% up to 82% was achieved. As a result of this, having a system that can classify a paragraph with a precision close to that of a human baseline, is seen as impressive.

Previous work completed in the Norwegian domain is work such as Hammer et al. [11], where sentiment lexicons were made. Research in the financial news domain by Njølstad et al. [18], is one this research builds its work upon. This paper is therefore an addition to the field of sentiment analysis on Norwegian texts.

3 System Overview

A high-level system description is adapted from Njølstad et al. [18], where an annotated data set is given as input. This data set is then used to extract Co-Occurring Terms (COTs) with the specified radius. These COTs are subsequently ranked using ranking functions described in Section 3.5 before they are manually annotated.

3.1 Co-Occurring Terms

Following [15] and [20], a Co-Occurring Term is whenever words in a sentence that bear importance to each other, co-occur. As our work is a test to see if the results in [18] can be replicated in a different domain, we will follow many of their definitions, as well as implementing their algorithms for COT generation and ranking, to be sure that our results can be as closely compared to theirs as possible. As a result of this, in this paper a COT can occur anywhere in a sentence within the paragraph. It is decided by two factors: *arity* - the number of terms the COT is composed of, set to a default of 2 in this paper, and *radius* - the maximum distance, measured in number of words, allowed between the two terms in the COT.

There are cases where a term can be comprised of two words to better comply with the Norwegian language. An example of such a case is a "*preposisjonsuttrykk*", which is defined in Store Norske Lekson (*Big Norwegian Lexicon*) as a preposition and a nominal clause, and serves as an adverbial or adjectival clause in a sentence [24]. Look at the COT ["skal", "i stedet"] ("*shall*", "*in place*"). Usually a COT would just have two words: ["skape", "arbeidsplasser"] ("*create*", "*working places*"), but it does not make much lexical sense to divide up "i stedet" as the two words will not serve the same purpose in the sentence if they are. As our POS tagger also groups these two words into one single term, we have decided to allow this behavior in our research.

The radius variable is significant as it is what makes COTs different from n-grams. Where an n-gram specifies that the two terms need to be adjacent (radius = 1), a COT can consist of two terms which are further apart from each other (radius > 1).

There has also been put a limitation on which words to look for in terms of creating a COT. Verbs, adverbs, nouns, and adjectives are the classes of words that in most cases bear a sentiment value [2]. We have used the Oslo-Bergen tagger[1] to identify the part-of-speech in the text.

3.2 Data

The data used for this paper was collected from two different, online news sources, NRK and VG, two of the biggest newspapers in Norway, over the course of 4 months (1st of July - 1st of October 2015). In total this amounted to 1108 news articles.

However, for the purpose of our system, annotation was done using 3961 paragraphs randomly picked from this data set. Two of the authors annotated all paragraphs as either positive (1), negative (-1) or neutral (0), ending up with two different sets of annotated data sets. This gave us a joint-probability of agreement of ∼76%. As we illustrated in Section 2, previous work on manual annotation tasks show agreement between humans in the range of ∼63% to ∼82%, which puts our human baseline in the same range as the research we compare our results with.

[1] http://www.tekstlab.uio.no/obt-ny/

3.3 Classifiers

Njølstad et al. [18] compared three different classifiers, namely, J48, Random Forest (RF), and Naïve Bayes (NB), for their analysis of financial news sentiments. For this paper, all three aforementioned are also being tested. Mainly to see if results can be replicated using same methods in a different domain, but also because these are simpler and less computationally costly than Artificial Neural Networks (ANN) and Support Vector Machines (SVM)[30]. In sentiment analysis applications, SVM have been widely used and researched, though J48 and RF have recently proved to produce higher precision in the financial news domain [18]. The machine learning classifiers used in our system are all set up using the WEKA framework[2].

3.4 Lexicon

To find the lexicon best suited for our domain, we test each machine learning classifier with all combinations of lexicon parameters. We have three different parameters: the size of the lexicon $\sigma = \{1000,2000,4000\}$, radius of the COTs $r = \{2,4,6,8\}$, and the ranking function $f = \{tf, idf, tfidf, mi, \chi^2\}$. The lexicon size, σ, is the maximum number of COTs allowed in a lexicon, and the radius, r, is the maximum distance between the terms in a COT as we mentioned in Section 3.1.

Njølstad et al. [18], argues that having a bigger lexicon size will, in general, increase the precision of the system. The same research also shows that this effect is diminishing as the size gets greater. Doubling the size from $\sigma = 1000$ to $\sigma = 2000$ gives a precision increase of $\sim$6%. Doubling again to $\sigma = 4000$ only gives an increase of $\sim$4%. Since it is imperative for our method that the time taken to manually annotate the lexicon is reasonable, we conclude that the additional work associated with lexica exceeding 4000 entries, does not justify the small gains in precision.

As these lexicon sizes are absolute, it is important to note that, in a few cases the actual size of the lexicon could be less than 4000. The reason for this is that smaller radii may not have as many as 4000 candidate COTs. This may lead to a misleading benefit of increasing the radius when $\sigma = 4000$ because the number of candidate COTs will be greater. Njølstad et al. deals with this by also using a lexicon size relative to the available candidate COTs given the radius. The absolute size is on the other hand given the most weight by the authors as one of their goals is to keep the size small so that the manual effort does not exceed that which is deemed reasonable. In addition, we are mostly interesting in comparing our results in a new domain, and thus absolute size is the only size we are using in this paper.

Just as we capped our lexicon size at 4000, we employed a maximum radius of $r = 8$. A greater radius will yield more COTs, however, at the same time the relationship between the two terms in a COT will be weaker as they are

[2] http://www.cs.waikato.ac.nz/ ml/index.html

further apart from each other. Again, the chance of benefiting from using a bigger radius does not outweigh the increased effort of doing so, and we are therefore only testing with radii from 2 to 8.

3.5 Ranking Functions

To make sure that the size σ of the lexicon is upheld, we need to limit the number of candidate COTs which will be included in our lexicon. To find the COTs that are significant to our domain, and exclude those that are not, we implemented the five ranking functions used in [18].

Term Frequency (TF) is used to measure how often a COT occurs at the collection level, Inverse Document Frequency (IDF) for measuring the specificity of a COT over the collection of documents, and TF-IDF that weights together the specificity of the COT, occurring infrequently across the collection. These are implemented for ranking the full set of candidate COTs, and then only pick the top number of COTs − those that have a document frequency greater than 1. In turn these will be annotated and used in the sentiment lexicon. For measuring the tendency of two terms that bear a meaning to each other, bias functions have been implemented, which are Mutual Information (MI) and Chi-Squared (χ^2). Both MI and χ^2 are used for measuring the degree of bias in a COT.

4 Results and Evaluation

The intent of this work is to acquire a sentiment lexicon by extracting COTs from a data set of paragraphs from Norwegian political news articles. The overall goal is to assess to what extent a successful sentiment analysis approach from finance can be applied in the more complex and subtle political domain. We have configured our system as closely as possible to that of [18] so that our results can be easily compared, with only smaller adjustments to accommodate for the difficulties in sentiment target selection in our domain. These adjustments should not have any significant effects on our results as we are still doing sentiment analysis on the document-level − treating each paragraph as a document.

Our system was tested with 180 unique combinations of parameters: lexicon size $\sigma = \{1000,2000,4000\}$, COT radius $r = \{2,4,6,8\}$, ranking function $f = \{tf, idf, tfidf, mi, \chi^2\}$, and machine learning classifier $c = \{nb, rf, j48\}$. The total amount of available candidate COTs were 9643, which means that many COTs appear in several of the different lexica. The manual effort of annotating COTs was an important factor in the work done in [18], where the authors spent ∼4 hours annotating their combined lexicon of 7990 COTs. They deemed this a feasible amount of manual labor, which lead us to do the same with the ∼5 hours we spent annotating COTs.

The results of running our system with all combinations of parameters ($r \times \sigma \times f \times c$) are presented in Table 1 with the highest precision of 59.7% highlighted in bold.

Table 1. Precision of system with input parameters ($r \times \sigma \times f \times c$)

		σ = 1000				σ = 2000				σ = 4000			
MLC	Ranking Function	COT radius = 2	4	6	8	COT radius = 2	4	6	8	COT radius = 2	4	6	8
NB	*tf*	54.9	55.1	54.9	54.0	56.8	57.3	56.3	55.1	53.6	57.4	57.9	56.6
	idf	52.4	52.5	52.0	51.4	50.8	53.6	52.6	52.9	53.6	51.6	51.3	51.7
	tfidf	55.3	55.2	54.6	53.9	56.8	57.5	56.8	54.9	53.7	57.2	57.9	56.6
	mi	51.9	51.3	50.8	50.9	52.3	52.5	52.1	52.8	53.6	53.5	53.8	52.9
	χ^2	53.3	52.1	52.2	50.8	53.5	52.8	53.1	52.4	53.6	54.3	52.6	52.8
RF	*tf*	51.3	51.1	51.3	48.6	52.6	52.3	52.0	50.7	50.2	54.5	53.1	53.2
	idf	49.9	48.1	48.2	49.1	47.8	49.2	47.9	48.2	50.3	47.5	47.2	48.2
	tfidf	52.2	51.5	51.3	48.5	52.8	52.5	52.4	50.7	50.3	54.2	53.3	52.5
	mi	48.8	48.9	47.2	47.3	47.8	50.3	48.8	47.7	50.3	49.6	49.1	49.0
	χ^2	47.2	49.4	48.6	48.0	49.1	47.8	47.1	48.8	50.3	50.5	46.7	47.1
J48	*tf*	56.0	55.7	55.9	54.3	57.4	58.0	56.3	56.2	54.8	59.0	59.5	58.8
	idf	51.6	51.5	51.1	50.6	50.3	52.5	51.4	51.8	54.8	50.0	51.8	52.0
	tfidf	55.6	56.0	55.4	54.5	57.5	57.8	56.6	56.1	54.7	**59.7**	59.5	57.7
	mi	51.2	51.1	49.3	51.3	53.7	52.7	53.0	52.5	54.8	53.6	54.7	53.1
	χ^2	53.4	51.8	51.0	49.8	53.3	53.3	52.9	51.9	54.7	53.2	52.5	54.3

Comparing the results with our own human baseline of ∼76%, we see that there is a substantial 16% difference in precision. Even though we cannot expect the machine learning classifier to outperform the work of a human being in the context of sentiment analysis, we would without doubt want to have precision results closer to our baseline. If we look at the related work we wanted to compare ourselves with, we are still pretty far off. Njølstad et al. achieved state-of-the-art precision results, with their highest at 69.1%. However, this was achieved with a relative lexicon size. Since we are only using absolute lexicon size, we should compare with the precision achieved with an absolute lexicon size in [18]. When doing so, we are getting closer as the highest result here is 65.9%. Looking at all the different combinations it is easy to see that our results lie ∼5-10% lower than the ones we are comparing with. What is most concerning here is that our maximum precision is closer to that of a guessing game than it is the state-of-the-art classification performance.

When the results are all so close to 50%, it can be hard to draw conclusions about the trends and the effects of each input parameter, as the closer you get to guessing, the results may seem random at first glance. A closer look at Table 1 reveals some of the same trends as were discovered in [18]. One of the hypotheses we stated earlier, was that COTs with smaller radii, i.e. with terms closer to each other within a sentence, have stronger relationships. If this were true then smaller radii would give better results. As we can see in our table, the results are in general declining as the radius gets bigger for each lexicon size which backs up this hypothesis. There are of course some exceptions to this, however, this could be a symptom of the coupling of radius and lexicon size when using an absolute size as mentioned in Section 3.4.

Looking at the general effect of increasing the lexicon size our results show that performance is somewhat better when lexicon size increases. The results seem to support our claim in Section 3.4 where we anticipated this trend. The difference in our results from [18] is that we do not see the exact same effect where the performance increase is declining each time the size is doubled. When the overall precision is so close to a guessing game, it makes sense that there should be more of a random distribution, and not as clear a trend as the results we are comparing with show.

When it comes to the ranking functions, the function *term frequency-inverse document frequency* ($tfidf$) scores the best. It is difficult to say why this is the case just by looking at the results we have here. However, going deeper into our results and looking at which sentiment classes are correctly classified and when they are correctly classified, and vice versa, we can try to answer questions about why the results are so different to other research.

For the machine learning classifier, c, J48 scores higher than the other two. The same is the case in [18]. The more surprising result here is that Random Forest achieves the worst results, as opposed to Naïve Bayes in [18]. It is hard to say why this is, but again, this could very well stem from the fact that the overall precision is so poor.

5 Discussion

Results in this paper suggests difficulties in terms of analyzing political text. Theorized reasons for this, also suggested by studies to that of [1] and [23], is the language used by journalists, as well as the textual structure of articles. Subtle ways of including sentiments, requiring world knowledge, and few occurrences of sentiment-bearing words, are proving difficult for a sentiment analysis system to capture.

Compared to [18], precision in general has been lower with the use of different classifiers and ranking functions. Results from our study had a wider range of precision, spanning as much as twelve percent, and rarely reaching closer to the 60% mark, compared to the equivalent study performed in the financial news domain in [18], where most results were in the 60% echelon.

Even though we are analyzing on the paragraph level, compared to the article level in [18], the lexicon acquisition method has in fact been done on the article level. We found that the document frequencies and the term frequencies of COTs were the same for most COTs when running the algorithms on the paragraph level. This lead to results that did not show any meaningful difference from one ranking function to another as all the ranking functions are based on these two parameters in some way. Since the only reason for doing our sentiment analysis on the paragraph level has to do with sentiment targets, this does not affect the training of the machine learning classifiers as we are still acquiring our lexicon in our domain of research. The results we show in this paper are thus from acquiring our lexica on the article level, which proved to give more differentiated results. Understanding why our results differ justifies the need for discussion.

The human baseline in both studies does not vary much either, both around the 70% mark. One can ask if there has been any mistakes during the annotation process. A thorough inspection of the annotated data set, however, reveals that the annotations are of the desired quality. In the process of manual annotation, both annotators followed an agreed-upon sample of the guidelines proposed by [17]. This gives reason to believe that any major error has not been made in the annotation process, either.

This turns focus to the language itself. It is described in [9] to be difficult to understand for a human without extensive prior knowledge in the domain, and the journalists use complex structures to convey sentiments [1]. An example of this is when they intentionally include sentiments in a sentence without the use of evidently positive or negative words. In turn, this results in a paragraph being ranked as neutral, while a human being understands the paragraph as bearing a particular sentiment. Take the paragraph,

> Venstre-representanten mener bestemt at filmen 'Trygghet i hverdagen', som kostet 50.000 kroner å lage, er en valgkampvideo for Frp, ikke en informasjonsvideo for Justisdepartementet.
> *The representative for the Liberal party firmly believe that the film "Security in everyday life", which cost 50,000 kroner to create, is a campaigning video for Frp, not an information video for the Department of Justice.*

which was annotated by both authors as having a negative sentiment. Reasoning for this is because of the understanding that using government money to create a campaigning video instead of an informational video, is seen as something negative. Use of an intensifier like "bestemt" (*firmly*) in addition to mentioning an actual cost, to portray the opinion of the Liberal party representative, is what will make a human see a sentiment. A system will not have this world knowledge unless explicitly told. The different COTs, ["mener", "bestemt"] (*"firmly", "believe"*), ["kostet", "kroner"] (*"cost", "kroner"*), ["ikke", "informasjonsvideo"] (*"not", "information video"*), can not be classified as either positive or negative, as the pair on its own does not bear a sentiment value. The COT, ["Trygghet", "hverdagen"] (*"Security", "everyday life"*) can even be seen as positive, although it was not picked as a COT for the given lexicon due to its low document frequency. As there can also often occur more than one target in a political news article, often shifting between the different ones, this can also offer up some problems. Though, after analyzing the data set used, more often than not, when several targets occur, and especially political parties, it is in general a summary of how the parties fare in regard to polls. It is therefore believed to not cause the mass-discrepancy shown in this study.

In Table 2 the confusion matrix from the best result is shown, where the first row indicates which class a paragraph was classified as, whereas the column on the right is the actual class the paragraph belongs to. Of the 3016 available candidate COTs, only 1800 are correctly classified (highlighted numbers). A larger majority of the false negatives (the portion of incorrectly classified paragraphs which lies in the same row as the actual class of interest) for class positive (1)

Table 2. Confusion Matrix for best result, 59.7%.

a	*b*	*c*	← Classified as
343	242	113	*a = 1*
178	**1054**	194	*b = 0*
129	360	**403**	*c = -1*

and negative (-1), 242 and 360 respectively, fall into the classification of neutral (more than twice as many). Less fall into that of its opposite sentiment. As previously noted, we hypothesize that paragraphs which can seem neutral to the computer system, does in fact have a sentiment, though hidden in the complex structure of the language − something the human reader can fathom, whereas the system have problems with this effect. The number of false negatives classified as neutral defends this. As our system is heavily based on the use of COTs, this problem gives us reason to believe that the replication of the approach from [18] is not sufficient to yield state-of-the-art precision results in our domain.

6 Conclusion and Future Work

In this paper we have replicated the methods from [18] and implemented a sentiment lexicon based on COTs in a new domain without any prior existing lexical tools. We hypothesized that our domain, the Norwegian political news domain, is more unstructured and semantically complex than the financial news domain, which was researched in [18]. These methods yielded lower precisions overall in our domain with a maximum precision of 59.7%, compared to the 65.9% in [18]. From this we conclude that a lexicon built solely on annotated COTs is not enough to achieve state-of-the-art classification precision. Further investigation of our results, and the language used in our domain is detailed in section 5, which leads us to believe that our hypothesis is upheld by our analysis.

As results compared to other domains are significantly worse, we have recognized characteristics that make the political news domain more difficult:

- Political text makes use of few sentiment-bearing words.
- Sentiments are created through contextual knowledge and are seldom expressed directly.
- The political news domain is unstructured.

Our investigation seems to suggest that political news journalists use complex sentence structures to convey sentiment, instead of polarized terms that can be caught by a sentiment lexicon. This is supported by the fact that there is a considerable amount of *false negatives* in the classes 1 and -1 (positive and negative paragraphs) that are wrongly classified as neutral. The ability of a human to understand such sentence structures and see the journalists' intent of conveying sentiment in an objective way, is not caught by our sentiment analysis system.

As we argue that a method that makes use of a sentiment lexicon based on COTs is not enough in our domain, we propose that it could still be viable with some extra features that take into account the complexity of the sentence structures. We would need to do an analysis of the paragraphs that are annotated as sentiment bearing, combined with the results from this research, and find patterns that could be useful in classifying paragraphs. One suggestion would be to look at the use of negations when conveying sentiment, and then do a binary classification before polarizing.

References

1. Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., Van Der Goot, E., Halkia, M., Pouliquen, B., Belyaeva, J.: Sentiment analysis in the news. arXiv preprint arXiv:1309.6202 (2013)
2. Benamara, F., Cesarano, C., Picariello, A., Recupero, D.R., Subrahmanian, V.S.: Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In: ICWSM. Citeseer (2007)
3. Blaz, F., Galleguillos, C., Cristianini, N.: Detecting the bias in media with statistical learning methods text mining: Theory and applications (2009)
4. Chung, J.E., Mustafaraj, E.: Can collective sentiment expressed on twitter predict political elections? In: AAAI. vol. 11, pp. 1770–1771 (2011)
5. Ethnologue languages of the world. `https://www.ethnologue.com/statistics/country`, accessed: 2016-01-23
6. Evgenia, B., van Der Goot, E.: News bias of online headlines across languages. The study of conflict between Russia and Georgia (2008)
7. Feldman, R.: Techniques and applications for sentiment analysis. Communications of the ACM 56(4), 82–89 (2013)
8. Ferguson, P., O'Hare, N., Davy, M., Bermingham, A., Sheridan, P., Gurrin, C., Smeaton, A.F.: Exploring the use of paragraph-level annotations for sentiment analysis of financial blogs (2009)
9. Grøttum, E.T., Aalberg, T.: De vanskelige nyhetene–hvordan krav om politiske forkunnskaper og kildebruk kan gjøre nyhetsdekningen vanskelig å forstå. Norsk statsvitenskapelig tidsskrift 28(01), 3–23 (2012)
10. Gulla, J.A., Auran, P.G., Risvik, K.M.: Linguistics in large-scale web search. In: Natural Language Processing and Information Systems, pp. 218–222. Springer (2002)
11. Hammer, H., Bai, A., Yazidi, A., Engelstad, P.: Building sentiment lexicons applying graph theory on information from three norwegian thesauruses. Norsk Informatikkonferanse (NIK) (2014)
12. Hsueh, P.Y., Melville, P., Sindhwani, V.: Data quality from crowdsourcing: a study of annotation selection criteria. In: Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing. pp. 27–35. Association for Computational Linguistics (2009)
13. Huberman, B.A., Adamic, L.A.: Internet: growth dynamics of the world-wide web. Nature 401(6749), 131–131 (1999)
14. Lu, Y., Castellanos, M., Dayal, U., Zhai, C.: Automatic construction of a context-aware sentiment lexicon: an optimization approach. In: Proceedings of the 20th international conference on World wide web. pp. 347–356. ACM (2011)

15. Matsuo, Y., Ishizuka, M.: Keyword extraction from a single document using word co-occurrence statistical information. International Journal on Artificial Intelligence Tools 13(01), 157–169 (2004)
16. Melville, P., Gryc, W., Lawrence, R.D.: Sentiment analysis of blogs by combining lexical knowledge with text classification. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 1275–1284. ACM (2009)
17. Mihalcea, R., Banea, C., Wiebe, J.M.: Learning multilingual subjective language via cross-lingual projections (2007)
18. Njølstad, P.C.S., Høysæter, L.S., Gulla, J.A.: Optimizing supervised sentiment lexicon acquisition: Selecting co-occurring terms to annotate for sentiment analysis of financial news
19. Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity. In: Proceedings of ACL. pp. 271–278 (2004)
20. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and trends in information retrieval 2(1-2), 1–135 (2008)
21. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. pp. 79–86. Association for Computational Linguistics (2002)
22. Perez-Rosas, V., Banea, C., Mihalcea, R.: Learning sentiment lexicons in spanish. In: LREC. vol. 12, p. 73 (2012)
23. Rozina, G., Karapetjana, I.: The use of language in political rhetoric: Linguistic manipulation. Süleyman Demirel Üniversitesi Fen-Edebiyat Fakültesi Sosyal Bilimler Dergisi 2009(19) (2009)
24. Simonsen, H.G.: preposisjonsuttrykk (2014), `https://snl.no/preposisjonsuttrykk`, online; May 11th, 2016
25. Smith, A.: The internet and campaign 2010. `http://www.pewinternet.org/2011/03/17/the-internet-and-campaign-2010/`, accessed: 2016-02-12
26. Tavakolifard, M., Gulla, J.A., Almeroth, K.C., Ingvaldesn, J.E., Nygreen, G., Berg, E.: Tailored news in the palm of your hand: a multi-perspective transparent approach to news recommendation. In: Proceedings of the 22nd international conference on World Wide Web companion. pp. 305–308. International World Wide Web Conferences Steering Committee (2013)
27. Turney, P.D.: Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th annual meeting on association for computational linguistics. pp. 417–424. Association for Computational Linguistics (2002)
28. Wang, H., Can, D., Kazemzadeh, A., Bar, F., Narayanan, S.: A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In: Proceedings of the ACL 2012 System Demonstrations. pp. 115–120. Association for Computational Linguistics (2012)
29. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the conference on human language technology and empirical methods in natural language processing. pp. 347–354. Association for Computational Linguistics (2005)
30. Zhao, Y., Zhang, Y.: Comparison of decision tree methods for finding active objects. Advances in Space Research 41(12), 1955–1959 (2008)

[This page is intentionally left blank.]

Chapter 7

Paper II

Patrik Fridberg Bakken, Terje André Bratlie, Jon Atle Gulla: *Understanding the Political News Domain - Analyzing Negation Count and Co-occurring Terms in Sentiment Analysis*, to be submitted to either the 16th IEEE International Conference on Data Mining (ICDM 2016), or the 26th International Conference on Computational Linguistics (COLING 2016).

Understanding the Political News Domain - Analyzing Negation Count And Co-occurring Terms in Sentiment Analysis

Patrik F. Bakken, Terje A. Bratlie and Jon Atle Gulla
Departement of Computer and Information Science
Norwegian University of Science and Technology
7491 Trondheim
Email: bakken.patrik@gmail.com, terje_ab@hotmail.com, jag@ntnu.no

***Abstract*—In the domain of political news for sentiment analysis there has been little research done in previous years. This domain has been deemed complex and unstructured, which makes it hard achieve state-of-the-art precision results. Understanding the characteristics of the political language and the complexity of the way journalists convey sentiment is thus imperative to improve performance. We take use of a two-step binary classification model to experiment with different feature combinations. This lets us pinpoint the areas of effect of each feature, and further start to understand the unique aspects of the domain in question. After combining features such as negations count and co-occurring terms, we find that negation count does not have any impact on precision. However, it is very interesting to note that by adapting a method using co-occurring terms into the political news domain, we see in increase in precision in the polarity classification step. Improvements should be focused on subjectivity detection in the future, which would greatly increase the overall precision of the sentiment engine.**

I. Introduction

Speaking of sentiment analysis, one usually distinguish between supervised and unsupervised approaches [1]. Within these approaches, one common technology is the use of classification tasks, where problems can be ranked according to their text, or classified into predefined classes. Of these, focus will be on subjectivity and polarity classification, namely binary classification. The former takes into account that not all incoming text is opinionated, and that a system might need to be able to distinguish between what is subjective and what is objective in the span of a text. The latter assumes text to be opinionated, and therefore classify the text as falling in under one of two sentiment polarities – in general, positive or negative [2]. The two are not converse. As Mihalcea et al. [3] suggests, improvements in the more difficult subjectivity detection, will have a positive impact on polarity classification.

News articles are in general written by journalists who want to appear objective, who will refrain from using subjective vocabulary themselves. If sentiment were to appear, it might be in the form of a quote of someone with the opinion of that of the journalist, facts being omitted, focus being shifted to facts that portray an opinion, etc. This complex structure is what makes news articles more difficult to analyze for sentiment than product reviews [4].

In natural language, negations usually shift the polarity of a sentence. This is not something that has been widely focused on in sentiment analysis, though by including this word class, is something that will improve the understanding of a sentence and phrase [5]. The occurrence of negations in a sentence will vary, but we theorize that in a negative sentiment-bearing paragraph, there will be more negations, leading us to believe that this can positively contribute to classify a paragraph's sentiment.

Research conducted in the Norwegian political news domain by [6] reports results not satisfactory for a state-of-the-art sentiment analysis system. Research in our paper build heavily upon said work, and will further include additional methods and features, as proposed.

Taking all of this into account, we want to implement a two-step binary classifier, with the inclusion of negation count, coupled with the previously used co-occurring term lexicon by [6]. As there are limited research done in the language of Norwegian on this matter, and similarly little research in the political news domain, no reasonable estimation of results can be made. We do, however, intend to improve on past research, by focusing on methods described in this paper, and also gain better

understanding of which step these methods have an impact on.

The remainder of this paper will focus on related work (Section II), a system overview with the proposed methods being implemented for the Norwegian political news domain (Section III), results and evaluation of results in Section IV, followed up by discussion and conclusion in Section V and VI, respectively.

II. Related Work

The more widely researched technology of classification, is that of polarity classification. Work such as [7]–[10] focus on distinguishing an author's polarity towards a certain topic or object.

Analyzing a span of text, and detecting subjectivity, is something that has been more researched recently, and especially work conducted by Yu et al. [11] focus on separating subjective texts from those that portray factual information.

A combination of these two classification methods, can be seen in work completed by Wilson et al.[5]. Here, a two-step binary classification task takes place, where in the first step, filtering out neutral expressions, combined with the second step, classifying the polarity of the expression. This has shown to give better results than that of baseline.

Wilson et al. [5] also have a focus on negations in their work. By only having a predefined lexicon with positive and negative words, Wilson et al. argues that a phrase's sentiment will not be correctly captured. Taking into account the contextual and prior polarity, gives reason for better understanding of the phrases. Similar work to that of Wilson et. al., has been completed by [9], [10], [12], however with more focus on local negation.

News articles have not been widely researched when it comes to sentiment analysis. Finding bias across different news sources, and initial efforts for sentiment analysis in the news, have been conducted by Evgenia et al. [13] and Blaz et al. [14], respectively.

Balahur et al. [15] states that news articles are much different than blogs or product reviews, mostly down to seemingly objective language, and that the sentiment analysis has to be rethought in order for satisfactory results in those domains to take place.

Similar work to that of ours, has been completed in the Turkish political news domain by Kaya et al. [16], using machine learning and natural language processing. Satisfactory results were achieved, with accuracies between 65% to 77%.

Sentiment scores below the 60% mark was achieved in [6], with the use of an annotated lexicon and machine-learning. It is said work that is further researched in this paper, with the implementation of similar methods described in this section.

III. System Overview

A. Classifiers

Previous work by Bakken et. al. in [6] experimented with three different classifiers in order to test which one would yield the best results with their methods. As we are continuing their work, we want to use the same three classifiers; J48, Random Forest (RF), and Naïve Bayes (NB). J48 shows the highest precision scores, and we expect to see the same in our own results. These classifiers were chosen in order to minimize the change of variables from previous work [cite here], for their fast computational speed [cite here], and for their ease of use – less tuning is needed than the more widely researched Support Vector Machines (SVM). In addition, newer research also prove J48 and RF to yield higher precision results in the financial news domain [cite here]. All three classifiers in our system are set up using the WEKA framework[1].

As mentioned in I, in this research we are doing binary classification in two separate steps. This means that we will be assessing each classifier for both steps separately. Hence, the subjectivity analysis may prove the highest precision results for a classifier different from the polarization analysis.

B. Data and Annotation

All our training and testing data come from Norwegian news sources NRK and VG over a span of four months during the summer of 2015. This time frame was chosen because of the municipal elections in October that same year. The media should be covered with political news stories in the timeframe leading up to an election and thus gives us enough training data to build our system. As we show in Table IV, the total number of paragraphs used for training is 3016. Our data set is actually comprised of 3961, however, during our annotation study we reached an agreement of $\sim$76% between the two annotators, which left us with 3016 paragraphs. This annotation agreement gives us a human baseline to compare our results with, which, from [6] and [17], is well within the range of other studies of sentiment analysis.

[1]http://www.cs.waikato.ac.nz/ ml/index.html

TABLE I
POSSIBLE INPUT FEATURES FOR MACHINE LEARNING CLASSIFIERS

Features	Description	Type
WordCount	Number of words in a paragraph	Discrete
PositiveCots	Number positive COTs in a paragraph	Discrete
NeutralCots	Number of neutral COTs in a paragraph	Discrete
NegativeCots	Number of neutral COTs in a paragraph	Discrete
Negations	Number of negation words in a paragraph	Discrete
IsComment	Whether this paragraph belongs to an article which is a "Comment", not an "Article"	Binary

It is important to note that this annotation was done prior to our work in [6], which used the same data for a ternary classification model, which means that we do not have a definite human baseline for binary classification as this is considered to be a simpler task. We do believe however, that the amount of paragraphs where annotator A marks as sentiment bearing (either a positive or negative classification) and annotator B marks as bearing the opposite sentiment is fairly small. Hence, we theorize that doing a new, binary, annotation study is unnecessary work as our agreement percentage would only differ marginally. The cases where one of the annotators marks a paragraph as objective (neutral) should already be caught by the ternary annotation study.

From this ternary annotation study we need to group the positive (+1) and negative (-1) classes together into one subjective class in order to complete the subjectivity classification step (Step I-III in Figure 1).

C. Features

In our experimentation with binary classification of sentiment, we want to explore which steps that benefit from including a Co-Occurring Terms (COTs) based lexicon and a negation count to be able to provide information on the effects of these features. In Table I we list all the features we use in our experimentation. In Section IV we show results from each classification step for various combinations of features. Note that "IsComment" is always present as this is a category of news articles – categorized by the news source – we have found to always imply sentiment, which obviously helps in deciding whether a paragraph is subjective or not.

A short introduction to COTs is necessary to explain the use of these features. Hence, Section III-C1 illustrates what our COTs based sentiment lexicon looks like, and briefly how they are generated. Our reasoning behind the choosing of negation counts as the feature to experiment with will be detailed in Section III-C2 below. As this is not a completely arbitrary choice, it deserves some attention.

TABLE II
EXCERPT FROM SENTIMENT LEXICON

Term 1	Term 2	Sentiment
ta	ordet	0
er	allerede	0
stor	glede	1
er	misfornyd	-1
skape	arbeidsplasser	1
kan	svekke	-1
skal	i stedet	0

1) Co-Occurring Terms: Co-Occurring Terms (COTs), as defined in [2], [18], are words with importance to each other which co-occur in a sentence. These COTs form the basis for a sentiment lexicon which is used to extract features in steps 1 and 5 in Figure 1. Details on how this lexicon is acquired is outside the scope of this paper, however, it is described in depth in [6], [17]. Here, we will point out that we are using a lexicon built with the parameters which yielded the best results in [6], namely 59.7% precision. Thus, our COTs are generated with two "words", and the distance between these "words" is 4. They are also ranked with a *term frequency - inverse document frequency* ranking function to limit the lexicon size to 4000 COTs.

Table II shows an excerpt from the lexicon used in this research. It shows pairs of terms which make up a COT with the sentiment attached to it. Looking closely, the last pair of terms is not actually a "pair". Clearly Term 2 is made up of two words: "i stedet". This is not a mistake by the authors, but rather an extension of the definition of a *term* to better comply with the Norwegian language. The two words "i" (*in*) and "stedet" (*place*) are grouped into one single term by the Norwegian POS tagger – the Oslo-Bergen-Tagger[2] – because of something called "*preposisjonsuttrykk*". This is a grammatical term which Store Norske Leksikon (*Big Norwegian Lexicon*) defines as a sentence clause consisting of a preposition and a nominal clause, and serves as either an adverbial clause or adjectival clause in a sentence [19]. To divide these two words does not make much lexical sense, which is why the POS tagger keeps them as a single term. Thus, we decided to allow this behavior as well.

[2]http://www.tekstlab.uio.no/obt-ny/

TABLE III
NEGATIONS WORDS IN THE NORWEGIAN LANGUAGE

Norwegian *bokml*	**Norwegian** *nynorsk*	**English**
ikke	ikkje	not
ei	ei	*a version of "not"*
nei	nei	no
aldri	aldri	never
neppe	neppe	hardly
ingen, inga, intet	ingen, inga, inkje	no (*adjective*)

TABLE IV
NEGATIONS BROKEN DOWN IN CLASSES AND PARAGRAPHS

Annotated class	**Paragraphs**	**Negations per paragraph**
Positive	698	0.29
Neutral	1426	0.23
Negative	892	0.53
Positive+Negative	1590	0.43
All	3016	0.33

2) Negations: Negations are words which change the polarity of words, phrases or sentences. As our analysis is on the paragraph level, cannot go into each phrase use the negation word to change its polarity. However, while annotating and reading political news articles we got a feeling that negations were used by journalists to subtly imply sentiment without using strong sentimental words. If this were true, it could very well be used as a means to improve precision in one or both steps of a two-step binary classification. Table III shows the words we consider to be negations in the Norwegian language and their English translation if one exists.

As mentioned we are on the paragraph level, and thus we need a simple way to find out how the negations effect the whole paragraph, not just a single sentence or phrase. Our analysis of negations in our data set of paragraphs is illustrated in Table IV. The difference in negations per paragraph is what we are most interested in as the higher it is, the better the chance of actually having an impact as a feature in sentiment classification. The biggest difference is between the negative and neutral classes. On average, each paragraph annotated as negative has 0.53 negations, compared to only 0.23 per neutral paragraph. As will be discussed in Section III-D, the two binary classification steps are subjectivity classification and polarity classification. None of these steps are differentiating between negative and neutral paragraphs. However, the latter one classifies into positive and neutral classes, and if we look at the difference between these two in our analysis table, it seems to be close to a magnitude of 2 (0.29 vs. 0.53). This difference is interesting and is worth a closer examination, and thus the negations feature is included in the experimentation of polarity classification.

Next, we can see that there is about the same difference between the neutral class and the combined class of positive and negative paragraphs. The latter, which includes all sentiment bearing paragraphs – and will be referred to as the subjective class from now on – has 0.43 negations per paragraph versus 0.23 for the former. Again, this is interesting as this would justify the experimentation with negations as a feature in the subjectivity classification step also.

Counting negation words is a very simple method, though it could still prove to be helpful in sentiment analysis of Norwegian political news. We will get back to this in our evaluation and discussion in sections IV and V.

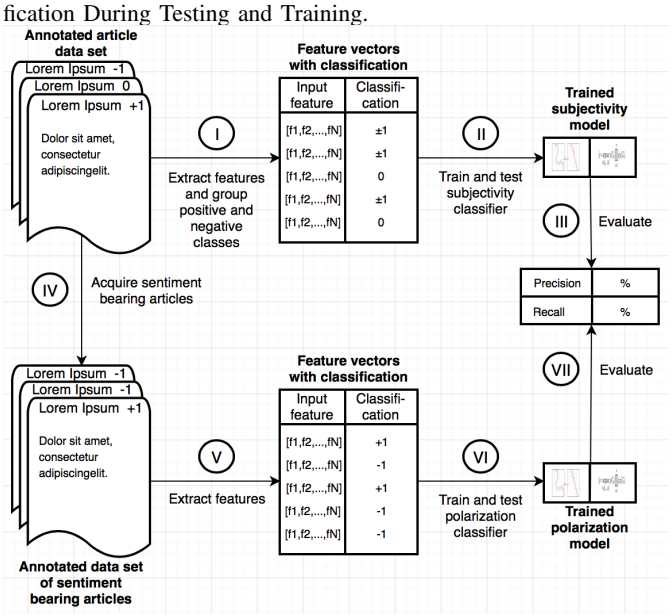
The main focus of this paper pertains to steps I, II, V and VI in Figure 1. The first two of these are related to subjectivity classification, whereas the last are part of the polarity classification. Subjectivity classification deals with the filtering of neutral (objective) paragraphs by classifying a paragraph as either objective or subjective. This is why all paragraphs annotated as negative or positive are grouped into one subjective class. With this classification step we can experiment with which features help determine whether a paragraph has sentiment or not.

D. High-level System Description

A two-step binary classification scheme requires two different machine learning models, just as the name suggests. For this reason we have implemented a system to train and test these models. Figure 1 illustrates a high-level overview of our system under the training and testing phases. The generation of COTs and acquisition of our sentiment lexicon is not depicted in this figure as this part of our system is outside the scope of this paper. If this part is of interest to the reader, we again refer to [6].

Polarity classification deals with the separation of positive and negative paragraphs. This classifier receives only subjective paragraphs, which means they all have to be either positive or negative. To train this machine learning model we have to remove all the neutral paragraphs from our data set so as to not "confuse" the model, which is shown in step IV. Because of this our training and testing system is not illustrated as a pipeline where the output of the subjectivity classification is the input of

Fig. 1. High-level System Overview With Two-step Binary Classification During Testing and Training.



the polarity classification as it would be in a live system. So, with this classifier we are interested in finding out which features that best determine whether a paragraph is negative or positive.

IV. Evaluation

A. Performance Evaluation

When evaluating sentiment analysis classifiers, there are a few different measurements to choose from. *Accuracy* is most commonly talked about, but not always in the correct way. It is included here only to point out the difference between the measurements. We are mostly interested in *precision* in our evaluation. However, when we looked at our results, we found it necessary to talk a little bit about the *recall* too. In this section we only include overall precision in the results to compare the different feature combinations. A more detailed discussion of some results comes in Section V.

Below is short description of the three most common measures and their respective equations ($N = tp + fp + tn + fn$):

- **Accuracy:** Fraction of instances classified correctly out of the total number of instances.

$$\frac{tp + tn}{N}$$

- **Precision:** Fraction of instances classified in the class of interest which really belong to this class.

$$\frac{tp}{tp + fp}$$

TABLE V
Subjectivity Classification, Precision Results With Various Feature Combinations.

	PosCots	NeutCots	NegCots	Negations	Precision
NB	✓	✓	✓		**67.3**
	✓	✓	✓	✓	66.4
				✓	59.9
RF	✓	✓	✓		62.8
	✓	✓	✓	✓	64.1
				✓	59.8
J48	✓	✓	✓		67.2
	✓	✓	✓	✓	67.0
				✓	61.8

- **Recall:** Fraction of instances actually belonging to the class of interest which are classified as such.

$$\frac{tp}{tp + fn}$$

B. Results

As we have mentioned already, the intent of this paper was to experiment with simple feature combinations in a two-step binary classification process, and achieve state-of-the-art precision in our domain of Norwegian political news. After our analysis of the use of negations in this domain we proposed the inclusion of a negation count as a feature. Bakken et. al. show precision scores below 60% by using a COTs based sentiment lexicon with a ternary classifier in the same domain [6]. Even though this is not at all close to the state-of-art in sentiment analysis, we have not ruled out the use of COTs as we still believe they could have a positive impact in a binary classification process. It is important to note that binary classification is a simpler task than ternary classification, and thus it can be challenging to compare directly with the work done in [6]. However, our experimentation will shed a light on which parts of the classification process that benefit from the various feature combinations shown in Table V and Table VI below.

Table V details our results from the subjectivity classification step. Highlighted in bold is the precision of the feature combination with the best result – 67.3%. Unsurprisingly this is ∼7% higher than the best results from [6]. As we can see from this combination, it does not include negations at all. We hypothesized that the difference in negations per paragraph between the neutral class and the subjective class would have a positive impact on the precision results for this step. Looking at these results, where there is a higher precision without the use of negations for both NB and J48, our hypothesis does not hold after all. RF is the only machine learning

TABLE VI
POLARITY CLASSIFICATION, PRECISION RESULTS WITH VARIOUS FEATURE COMBINATIONS.

	PosCots	NeutCots	NegCots	Negations	Precision
NB	✓	✓	✓		72.7
	✓	✓	✓	✓	68.1
	✓		✓		**72.9**
	✓		✓	✓	69.0
				✓	57.4
RF	✓	✓	✓		64.5
	✓	✓	✓	✓	64.4
	✓		✓		66.2
	✓		✓	✓	65.0
				✓	55.6
J48	✓	✓	✓		72.0
	✓	✓	✓	✓	71.9
	✓		✓		72.4
	✓		✓	✓	72.0
				✓	—

model which benefit from negations, though this model yielded unsatisfying results in general. One thing we can say for certain is that throwing out COTs all together performs the worst in every situation.

Looking at Table VI we have a few more combinations for the polarity classification step than we did for the subjectivity classification step. Here we are only differentiating between positive and negative paragraphs, and thus we could experiment with combinations that did not include neutral COTs. As we can see from the highlighted precision score, this is in fact what yields the best result of 72.9%, which is in line with state-of-the-art classification research. It is interesting again that our hypothesis that negations would have a positive impact on the results does not hold in this step either. In fact, even here, the use of negations performs at best at the same level as the combinations with all COTs features included. Without these features, we see the same as we did in step 1, which is that throwing out COTs yields very low precision scores.

For both steps NB is the machine learning model that performs the best. However, J48 outperforms the other two when including negations in the feature combination. Note that there is no score for J48 with only negations included. This is because this model was not able to build a decision tree based on this feature. RF on the other hand is again the worst performer.

There are some interesting results not shown in these two tables, which we discuss in the next section by going into more detail.

TABLE VII
STEP I: PRECISION AND RECALL FOR SUBJECTIVITY CLASSIFICATION

Class	Precision	Recall
0	57.4%	84.8%
± 1	**76.1%**	43.5%
Overall	67.3%	63.0%

V. DISCUSSION

When observing previous results from Bakken et al. [6], an overall precision of 59.7% was achieved. In our research, a two-step binary classification process has been employed, resulting in different precisions for both subjectivity classification (Step I), and polarity classification (Step II). Taken from Table VII, an overall precision of 67.2% was achieved for Step I. Similarly, 72.4% for Step II (Table VIII).

From Step I, the subjective class achieves a precision of 76.1%, which means that in a live system where the input of Step II is the output of Step I, 76.1% of the paragraphs analyzed in Step II will be correctly classified, which will better results. On the other hand, the amount of subjective paragraphs piped through to Step II, will be less due to the low recall value. As discussed in III-C2, an average of 0.43 negations per *Positive+Negative* paragraph, compared to 0.23 for *Neutral*, was theorized to have an impact on Step I. There were, with the inclusion of negation count as a feature, no noticeable higher gain in precision, leading us to believe that a difference of 0.2 negations on average per paragraph is not significantly high enough to achieve better results.

Again, precision in Step II is higher compared to that of Bakken et al. [6], with an overall precision of 72.9%. The difference of 0.24 negations more on average per *Negative* paragraph, compared to *Positive*, did not have an impact on polarity classification either, backing up the results of negation inclusion in Step I. The result in precision of 72.9% is from an experiment with the exclusion of neutral COTs, which makes the classifier focus more on the negative and positive COTs in a paragraph. This is something that can only work in classifying polarity, as subjectivity classification relies on neutrality, therein COTs.

In Step II, both the positive and negative class show interesting results, in terms of precision and recall respectively. Even if having an overall precision of 72.9%, and equally an overall recall of 72.1%, it is the characteristics of each class we would like to shift focus to.

TABLE VIII
STEP II: PRECISION AND RECALL FOR POLARITY CLASSIFICATION

Class	Precision	Recall
1	**76.1%**	53.3%
- 1	70.4%	**86.9%**
Overall	72.9%	72.1%

The *positive* (1) class shows meagre results in terms of recall, though in terms of precision ranks higher than the overall, achieving satisfactory results. For the *negative* (-1) class, precision is good enough, though it is its recall that is the impressive part, with a recall value of 86.9%. What both of these numbers tell us, is that there might be usage areas where this can be exploited. With a precision of 76.1% for positive sentiment bearing paragraphs, and ultimately articles, there might be systems that are more interested in having not all positive articles returned, rather can with a higher certainty say that its returned articles are positive. On the other hand, there might be systems that are not all too interested in having a high precision for saying a paragraph (article) is negative, rather want users to have access to most of the articles.

We would like to note, that binary classification is a simpler method than to include a third class, e.g. neutral, but ultimately gives better results in terms of precision, also emphasized by Wilson et al. [5]. By separating subjectivity classification into its own case, we achieve a higher precision, and as suggested by Mihalcea et al. [3], making improvements on subjectivity detection is something that ought to have a positive impact on polarity classification. Similarly, as suggested by [15], making improvements in the domain of news, is something that can further boost results. With the numbers achieved in this study, and possible usage areas that it can be applied to, it gives reason for further research.

VI. CONCLUSION AND FUTURE WORK

In this paper we have taken the research completed by Bakken et al. [6] further, by employing a two-step binary classification process. By doing so, close to state-of-the-art precision was achieved. We also experimented with the inclusion of negation count, as was theorized this would improve results in either subjectivity or polarity classification. No noticeable gain in precision was made, leading us to believe that the difference in negation count per polarized paragraph is not sufficiently high enough to distinguish the sentiment of paragraphs. As this was a simple method, not one where analysis of the placement of negation in terms of sentence structure took place, we do theorize that methods such as that of Wilson et al. [5], will in general achieve better results.

No similar studies that take use of a COT lexicon in a news domain, with a two-step binary classification process, has been found. As the outcome of the experiments indicates, with the exclusion of neutral COTs in the second step, polarity classification, we do achieve better results. Not significantly, though high enough to warrant a mention.

As our results show, distinguishing between two different steps when classifying polarity, achieves better results. Also backed up by Mihalcea et al. [3], that improvements on subjectivity detection is something that will ultimately improve polarity classification.

Research conducted in this paper, expanded from Bakken et al. [6], take use of paragraphs, and analyze these on a document-level. If distinguishing between sentences within a paragraph while similarly observing the negation count in each sentence, more specificity could be achieved.

All experiments have been conducted in the Norwegian political domain, and we therefore suggest that testing of proposed methods in similar domains, but different language, is worth exploring.

REFERENCES

[1] R. Feldman, "Techniques and applications for sentiment analysis," *Communications of the ACM*, vol. 56, no. 4, pp. 82–89, 2013.

[2] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

[3] R. Mihalcea, C. Banea, and J. M. Wiebe, "Learning multilingual subjective language via cross-lingual projections," 2007.

[4] A. Balahur, R. Steinberger, M. Kabadjov, V. Zavarella, E. Van Der Goot, M. Halkia, B. Pouliquen, and J. Belyaeva, "Sentiment analysis in the news," *arXiv preprint arXiv:1309.6202*, 2013.

[5] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics, 2005, pp. 347–354.

[6] P. F. Bakken, T. A. Bratlie, and J. A. Gulla, "Sentiment lexicon acquisition for sentiment analysis: Selecting co-occurring terms in norwegian political news," apr 2016, awaiting selection of conference to publish to.

[7] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 417–424.

[8] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods*

in natural language processing-Volume 10. Association for Computational Linguistics, 2002, pp. 79–86.

[9] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 168–177.

[10] S.-M. Kim and E. Hovy, "Determining the sentiment of opinions," in *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, 2004, p. 1367.

[11] H. Yu and V. Hatzivassiloglou, "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences," in *Proceedings of the 2003 conference on Empirical methods in natural language processing*. Association for Computational Linguistics, 2003, pp. 129–136.

[12] G. Grefenstette, Y. Qu, J. G. Shanahan, and D. A. Evans, "Coupling niche browsers and affect analysis for an opinion mining application," in *Coupling approaches, coupling media and coupling languages for information retrieval*. Le Centre de Hautes Etudes Internationales d'Informatique Documentaire, 2004, pp. 186–194.

[13] B. Evgenia and E. van Der Goot, "News bias of online headlines across languages," *The study of conflict between Russia and Georgia*, 2008.

[14] F. Blaz, C. Galleguillos, and N. Cristianini, "Detecting the bias in media with statistical learning methods text mining: Theory and applications," 2009.

[15] A. Balahur and R. Steinberger, "Rethinking sentiment analysis in the news: from theory to practice and back," *Proceeding of WOMSA*, vol. 9, 2009.

[16] M. Kaya, G. Fidan, and I. H. Toroslu, "Sentiment analysis of turkish political news," in *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01*. IEEE Computer Society, 2012, pp. 174–180.

[17] P.-C. S. Njølstad, L. S. Høysæter, and J. A. Gulla, "Optimizing supervised sentiment lexicon acquisition: Selecting co-occurring terms to annotate for sentiment analysis of financial news."

[18] Y. Matsuo and M. Ishizuka, "Keyword extraction from a single document using word co-occurrence statistical information," *International Journal on Artificial Intelligence Tools*, vol. 13, no. 01, pp. 157–169, 2004.

[19] H. G. Simonsen, "preposisjonsuttrykk," 2014, online; May 11th, 2016. [Online]. Available: https://snl.no/preposisjonsuttrykk

[This page is intentionally left blank.]

Bibliography

[1] Jon Espen Ingvaldsen and Jon Atle Gulla. Taming news streams with linked data. In *Research Challenges in Information Science (RCIS), 2015 IEEE 9th International Conference on*, pages 536–537. IEEE, 2015.

[2] Steve Lawrence and C Lee Giles. Accessibility of information on the web. *Nature*, 400(6740):107–107, 1999.

[3] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Internet: Diameter of the world-wide web. *Nature*, 401(6749):130–131, 1999.

[4] Bernardo A Huberman and Lada A Adamic. Internet: growth dynamics of the world-wide web. *Nature*, 401(6749):131–131, 1999.

[5] P. Simon. *Too Big to Ignore: The Business Case for Big Data*. Wiley and SAS Business Series. Wiley, 2013.

[6] Josh James. Data never sleeps 2.0. URL=https://www.domo.com/blog/2014/04/data-never-sleeps-2-0/, apr 2014. Accessed: 20.05.2016.

[7] Pew Research Center. Digital: Top 50 online news entities (2015). URL=http://www.journalism.org/media-indicators/digital-top-50-online-news-entities-2015/, jan 2015. Accessed 10.05.2016.

[8] Aaron Smith. The internet and campaign 2010. `http://www.pewinternet.org/2011/03/17/the-internet-and-campaign-2010/`. Accessed: 2016-02-12.

[9] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity. In *Proceedings of ACL*, pages 271–278, 2004.

[10] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

[11] Belyaeva Evgenia and Erik van Der Goot. News bias of online headlines across languages. *The study of conflict between Russia and Georgia*, 2008.

[12] Fortuna Blaz, Carolina Galleguillos, and Nello Cristianini. Detecting the bias in media with statistical learning methods text mining: Theory and applications, 2009.

[13] Maxwell E McCombs and Donald L Shaw. The agenda-setting function of mass media. *Public opinion quarterly*, 36(2):176–187, 1972.

[14] Maxwell McCombs. The agenda-setting role of the mass media in the shaping of public opinion. In *Mass Media Economics 2002 Conference, London School of Economics: http://sticerd. lse. ac. uk/dps/extra/McCombs. pdf*, 2002.

[15] Maxwell McCombs. *Setting the agenda: The mass media and public opinion*. John Wiley & Sons, 2013.

[16] Allan Bell. *The language of news media*. Blackwell Oxford, 1991.

[17] Ronen Feldman. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89, 2013.

[18] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

[19] Peter D Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics, 2002.

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

[21] Alexandra Balahur, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik Van Der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva. Sentiment analysis in the news. *arXiv preprint arXiv:1309.6202*, 2013.

[22] Farah Benamara, Carmine Cesarano, Antonio Picariello, Diego Reforgiato Recupero, and Venkatramana S Subrahmanian. Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In *ICWSM*. Citeseer, 2007.

[23] Yue Lu, Malu Castellanos, Umeshwar Dayal, and ChengXiang Zhai. Automatic construction of a context-aware sentiment lexicon: an optimization approach. In *Proceedings of the 20th international conference on World wide web*, pages 347–356. ACM, 2011.

[24] Hugo Lewi Hammer, Per Erik Solberg, and Lilja Øvrelid. Sentiment classification of online political discussions: a comparison of a word-based and dependency-based method. Association for Computational Linguistics, 2014.

[25] Hugo Hammer, Aleksander Bai, Anis Yazidi, and Paal Engelstad. Building sentiment lexicons applying graph theory on information from three norwegian thesauruses. *Norsk Informatikkonferanse (NIK)*, 2014.

[26] Aleksander Bai, Hugo Hammer, Anis Yazidi, and Paal Engelstad. Constructing sentiment lexicons in norwegian from a large text corpus. In *Computational Science and Engineering (CSE), 2014 IEEE 17th International Conference on*, pages 231–237. IEEE, 2014.

[27] Pal Christian S Njolstad, Lars S Hoysaeter, Wei Wei, and Jon Atle Gulla. Evaluating feature sets and classifiers for sentiment analysis of financial news. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on*, volume 2, pages 71–78. IEEE, 2014.

[28] Ethnologue languages of the world. `https://www.ethnologue.com/statistics/country`. Accessed: 2016-01-23.

[29] Mozhgan Tavakolifard, Jon Atle Gulla, Kevin C Almeroth, Jon Espen Ingvaldesn, Gaute Nygreen, and Erik Berg. Tailored news in the palm of your hand: a multi-perspective transparent approach to news recommendation. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 305–308. International World Wide Web Conferences Steering Committee, 2013.

[30] Jon Atle Gulla, Per Gunnar Auran, and Knut Magne Risvik. Linguistics in large-scale web search. In *Natural Language Processing and Information Systems*, pages 218–222. Springer, 2002.

[31] Pål-Christian Salvesen Njølstad and Lars Smøras Høysæter. Sentiment analysis for financial applications. Master's thesis, Norwegian University of Science and Technology, 2014.

[32] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics, 2005.

[33] Pål-Christian Salvesen Njølstad, Lars Smøras Høysæter, and Jon Atle Gulla. Optimizing supervised sentiment lexicon acquisition: Selecting co-occurring terms to annotate for sentiment analysis of financial news.

[34] Yanxia Zhang and Yongheng Zhao. Classification in multidimensional parameter space: Methods and examples. *Publications of the Astronomical Society of the Pacific*, 115(810):1006, 2003.

[35] Yongheng Zhao and Yanxia Zhang. Comparison of decision tree methods for finding active objects. *Advances in Space Research*, 41(12):1955–1959, 2008.

[36] J Ross Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.

[37] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[38] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):587–588, 2005.

[39] Kevin P Murphy. Naive bayes classifiers. *University of British Columbia*, 2006.

[40] Jason D Rennie, Lawrence Shih, Jaime Teevan, David R Karger, et al. Tackling the poor assumptions of naive bayes text classifiers. In *ICML*, volume 3, pages 616–623. Washington DC), 2003.

[41] Alexandra Balahur and Ralf Steinberger. Rethinking sentiment analysis in the news: from theory to practice and back. *Proceeding of WOMSA*, 9, 2009.

[42] Rada Mihalcea, Carmen Banea, and Janyce M Wiebe. Learning multilingual subjective language via cross-lingual projections. 2007.

[43] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145, 1995.

[44] Ralph Grishman. *Computational linguistics: an introduction*. Cambridge University Press, 1986.

[45] Roland Hausser and R Hausser. *Foundations of computational linguistics*. Springer, 1999.

[46] Paul R Kroeger. *Analyzing grammar: An introduction*. Cambridge University Press, 2005.

[47] Ordnett. `https://www.ordnett.no/spr%C3%A5kverkt%C3%B8y/spr%C3%A5kvett.ordklassene`. Accessed: 2016-04-19.

[48] Viktor Pekar. Linguistic preprocessing for distributional classification of words. In *Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries*, pages 15–21. Association for Computational Linguistics, 2004.

[49] Emma Haddi, Xiaohui Liu, and Yong Shi. The role of text pre-processing in sentiment analysis. *Procedia Computer Science*, 17:26–32, 2013.

[50] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schuetze. Introduction to information retrieval. In *DAGLIB*, pages 117–119, 271–279, 2008.

[51] Livia Polanyi and Annie Zaenen. Contextual valence shifters. In *Computing attitude and affect in text: Theory and applications*, pages 1–10. Springer, 2006.

[52] Alistair Kennedy and Diana Inkpen. Sentiment classification of movie reviews using contextual valence shifters. *Computational intelligence*, 22(2):110–125, 2006.

[53] Yutaka Matsuo and Mitsuru Ishizuka. Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, 13(01):157–169, 2004.

[54] Hanne Gram Simonsen. preposisjonsuttrykk, 2014. Online; May 11th, 2016.

[55] David Crystal. *The Cambridge encyclopedia of language*, volume 2. Cambridge Univ Press.

[56] John Daintith. A dictionary of computing, 2004. Online; May 20th, 2016.

[57] John S Uebersax. Diversity of decision-making models and the measurement of interrater agreement. *Psychological bulletin*, 101(1):140, 1987.

[58] Gunta Rozina and Indra Karapetjana. The use of language in political rhetoric: Linguistic manipulation. *Süleyman Demirel Üniversitesi Fen-Edebiyat Fakültesi Sosyal Bilimler Dergisi*, 2009(19), 2009.

[59] Ruth Wodak. *Language, power and ideology: Studies in political discourse*, volume 7. John Benjamins Publishing, 1989.

[60] Murray Edelman. *Political language: Words that succeed and policies that fail*. Elsevier, 2013.

[61] Eva-Therese Grøttum and Toril Aalberg. De vanskelige nyhetene–hvordan krav om politiske forkunnskaper og kildebruk kan gjøre nyhetsdekningen vanskelig å forstå. *Norsk statsvitenskapelig tidsskrift*, 28(01):3–23, 2012.

[62] Jessica Elan Chung and Eni Mustafaraj. Can collective sentiment expressed on twitter predict political elections? In *AAAI*, volume 11, pages 1770–1771, 2011.

[63] Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, pages 115–120. Association for Computational Linguistics, 2012.

[64] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10:178–185, 2010.

[65] Erik Tjong Kim Sang and Johan Bos. Predicting the 2011 dutch senate election results with twitter. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, pages 53–60. Association for Computational Linguistics, 2012.

[66] Murphy Choy, Michelle LF Cheong, Ma Nang Laik, and Koo Ping Shung. A sentiment analysis of singapore presidential election 2011 using twitter data with census correction. *arXiv preprint arXiv:1108.5520*, 2011.

[67] Adam Bermingham and Alan F Smeaton. On using twitter to monitor political sentiment and predict election results. 2011.

[68] Daniel Gayo Avello, Panagiotis T Metaxas, and Eni Mustafaraj. Limits of electoral predictions using twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. Association for the Advancement of Artificial Intelligence, 2011.

[69] Ahmed Mourad and Kareem Darwish. Subjectivity and sentiment analysis of modern standard arabic and arabic microblogs. In *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 55–64, 2013.

[70] Mesut Kaya, Guven Fidan, and Ismail H Toroslu. Sentiment analysis of turkish political news. In *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 174–180. IEEE Computer Society, 2012.