Experiments on Satire Detection for Norwegian News Articles

Katarina Ekren

Supervisor: Samia Touileb

Master's Thesis
Department of Information Science and Media Studies
University of Bergen

May 31, 2022

Scientific environment

This study is carried out at the Department of Information Science and Media Studies at the University of Bergen with the support of Web64. Web64 is a company collecting everything published publicly online in a country, and using AI to extract metadata. The Norwegian news articles used in this study are provided by Web64.

Acknowledgements

Samia Touileb I especially want to acknowledge my supervisor, Samia Touileb. Your knowledge and expertise have been invaluable. Thank you.

Olav Hjertaker I would also like to express my gratitude to Olav Hjertaker from Web 64, for introducing me to the satire detection project and for collecting news article datasets used in the study.

Fellow students Thank you to my fellow students at room 635 for being the best group of people to spend my study time with. For your motivation, laughter and support. You have been absolutely amazing.

Family and Friends I thank my family and friends for all your support, and for believing in me. It has been extremely appreciated.

Katarina Ekren
Bergen, 31.05.22

Abstract

The spread and amount of misinformation is increasing. The World Economic Forum (WEF) has listed it as one of the main threats to our society (*Howell*, 2013). Satire is one of the problems when it comes to misinformation, more specifically news satire. News satire is a genre of satire that resembles the characteristics of true journalistic reporting, while keeping the main objective of satire that is: use of a combination of humor and irony, usually with exaggeration, to expose and make fun of political or newsworthy issues.

In this thesis we present to the best of our knowledge, the first attempt to automatically detecting satire in Norwegian news articles. Automatically identifying satirical news pieces can aid in minimizing the potential deceptive impact of satire. To this end, we employ three classification methods, namely Naïve Bayes, SVM (support-vector machines) and logistic regression, based on TF-IDF (Term Frequency Inverse Document Frequency) feature weights. All three machine learning models achieved similar results.

In total, our dataset incorporates 6322 articles containing a balanced collection of satirical and non-satirical news texts from various domains (3161 satirical and 3161 non-satirical). Using this corpus we proposed three cross-domain satire detection tasks, one considering only the use of headlines, one considering only the use of article texts, and lastly one considering full articles, including both headlines and texts.

After experimenting on the test sets, we achieved the top accuracy score of 98% using only text as input, and the combination of title and text as input, with SVM. We observe that satire detection on news headlines was significantly more challenging, the top accuracy score being 76% using SVM. We believe that this shows that the automatic detection of satire using only headlines is quite challenging. Especially when using simple machine learning approaches, and we believe that this might be due to the length of headlines and the need for more context.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

"Wherever there is objective truth, there is satire"

Wyndham Lewis (1882-1957)

The spread and amount of misinformation is increasing. The World Economic Forum (WEF) has listed it as one of the main threats to our society (*Howell*, 2013). Since our world is now hyper-connected, this enables an escalated spread of information online. Deceptive and misleading news articles have been around for a while, and allows for either intentionally or unintentionally spread of misleading or provocative information. Which can lead to serious consequences, *i.e.,* the assault of the US Congress on 6^{th} of January 2021. The importance of the detection of deceptive news is therefore rising rapidly, as more and more people start relying on online news as their major source of information (*De Sarkar et al.*, 2018).

Defining misinformation can be challenging. A definition by *Nyhan and Reifler* (2010) distinguishes between misinformation (regarding the information itself) and misperceptions (the beliefs that people hold). It is important to clarify; that "while misinformation often generates misperceptions, the two are conceptually separate" (*Vraga and Bode*, 2020). *Nyhan and Reifler* (2010) also define misperceptions as "cases in which peoples beliefs about factual matters are not supported by clear evidence and expert opinion - a definition that includes both false and unsubstantiated belief about the world". According to *Shin et al.* (2018), misinformation gains its power when it is repeated and passed along from one person to another.

Social media sites, such as Facebook and Twitter are said to be platforms where fake news widely circulates (*Jiang et al.*, 2021). In nature, people generally cling to like minded, resulting in the creation of echo chambers (where similar minded people mimic each other's beliefs). The challenge with echo chambers is that information can be amplified and repeated within closed networks, causing ideas and beliefs to become

more aggressive (*Farrell*, 2015), thus increasing the spread of misinformation.

Satire is one of the problems when it comes to misinformation, more specifically news satire. News satire is a genre of satire that resembles the characteristics of true journalistic reporting. The satire articles are often inspired by real news, and cover the same range of topics from sports, to politics, to crime. It is essential to mention that satire differs from "fake" news, in the sense that fake news purposely try to deceive people by giving them untrue facts, while satirical news on the other hand, aims to ridicule and criticize something or someone, by producing satirical comments through fictionalized stories. Unlike fake news, where the objective is to persuade the readers to believe the news are true, satire has the author's intention of being recognized as "fake" (*De Sarkar et al.*, 2018).

Satirical news detection is important in order to prevent the spread of misinformation over the Web. In this study, the goal is to experiment and present a first attempt at detecting Norwegian satire news, presenting machine learning-based models to distinguish between satirical and non-satirical news content.

Norwegian is considered a minor language with limited textual resources when it comes to constructing huge corpora. Therefore, it has been difficult to train high-performing transformer-based models for such languages so far (*Kummervold et al.*, 2021).

1.1 Problem Statement

Due to the amount of misleading news on the Internet, there is a need for new techniques for preventing the spread of misleading news. Since satirical news articles are at least part of the time deceptive, identifying satirical news pieces can assist in minimizing the potential misleading effect of satirical news. Looking at the satire detection phenomena, there already exist working machine learning (ML) models to detect satire. However, to the best of our knowledge, this is the first and only dataset for the study of Norwegian satirical news. In this thesis, we will focus on detecting satire in Norwegian news articles. Creating and examining three simple ML models, and help contribute to further research.

1.2 Objectives and contributions

The main goal of this study is to try answering if it is possible to predict if an article is satire or not, based on textual content alone (not including pictures or illustrations). Another important contribution of this work is that we only will look at Norwegian

articles, collected from Norwegian news sources.

To reach our objectives, we focus on three research questions:

- **RQ1**: How will simple machine learning models perform in satire detection for Norwegian?
- **RQ2**: Which types of input achieve the highest classification scores?
- **RQ3**: Which aspects of satire are difficult to handle by simple machine learning models?

To solve the first Research Question (RQs1), we will use three simple ML models for text classification, namely Naïve Bayes, logistic regression, and a linear SVM. We have decided to use these models because they are simple, and therefore easier to understand their inner workings. This will allow us to better understand the most informative features during each classification task, and have a better overview of the types of errors they make.

For RQ2, we will be testing various types of input, and see how this affects the performances of our models. It has been shown that headlines play an important role in satire detection (*Burfoot and Baldwin* (2009) and *Rubin et al.* (2016)), and we aim to investigate if this applies to Norwegian and our dataset, but also which other inputs can yield good results. To this end, we will perform three tasks:

1. Task 1: use only title as input.
2. Task 2: use only text as input.
3. Task 3: use a combination of title and text as input.

We believe that this will shed light on the complexity of the task, and hopefully give us insights into how much contextual information is necessary for the task of satire detection.

For RQ3, we will perform an error analysis on the final classifications of our models on the test set, and explore which types of errors seem to be predominant. The main focus will be on understanding the types or errors and difficulties met by our simple ML models, which we believe can shed light on how to better approach the task of satire detection in the Norwegian language.

Despite our work focusing on simple machine learning models, we approach the problem of satire detection in an exploratory way. We employ simple models, give them various types of inputs, and analyse their outputs and types of errors they produce. Our main contributions are therefore:

- A collection of a Norwegian corpus, with an equal distribution of satirical and non-satirical articles from various domains.
- ML models for satire classification on the Norwegian language.
- An analysis of the performance of the ML models with various inputs.

1.3 Thesis outline

Chapter 2: In this chapter we give an overview of previous work carried out on English and other languages. We introduce the definitions of satire, news satire and related work on the topic of satire detection.

Chapter 3: Here we present our dataset, how we pre-processed it, and outline our experimental setups by describing our ML models, and how we approached the task of satire detection.

Chapter 4: We present our results for all three models, and all three tasks, both on the validation and test sets. We provide a discussion about our results, and do an extensive errors analysis. We also summarize our main findings, and discuss the limitations of our work.

Chapter 5: Here we present our conclusion as well as a summarization of our answers and solutions to our RQs. We also discuss our contributions, and further work that needs to be done on the topic of satire classification for Norwegian.

Chapter 2

Background

The rapid increase of data has caused an accelerated growth in the demand for text analysis techniques. These techniques extract information from textual data, like social media feeds, blogs, emails, news articles, and other forms of text sharing (*Gandomi and Haider*, 2015). This chapter will provide an overview of relevant approaches for trying to solve satire detection in Norwegian articles. First, in section 2.1 we give an introduction to satire, satire in news and satire detection. Then, section 2.2 gives an introduction to previous research on satire detection for English and other languages. Lastly, in section 2.4, we give a brief introduction of which text classification algorithms we will deploy in this thesis based on previous work.

2.1 Satire

Thrall and Hibbard (1960) defines satire as a "literary manner which blends a critical attitude with humor and wit to the end that human institutions or humanity may be improved". It is a way of criticizing a person, an idea or an institution in which you use humor to show their faults and weaknesses. However, if the audience does not understand the real intentions hidden in the ironic dimension, satire loses its significance (*Barbieri et al.*, 2015a). *Gilmore* (2017) argues that satire is only effective if it's perceived by persons other than its author to be such, and responses can change depending on factors such as time and circumstances.

Satire is traditionally divided into two main styles; *Horatian* and *Juvenalian*. The Horation style is the more playful of the two, whereas Juvenalian is more overtly hostile (*Rubin et al.*, 2016). Horation satire is characterized by mocking and dark humor. It aims to correct by soft and widely amenable laughter (*Thrall and Hibbard*, 1960). Juvenalian, on the other hand, is described as being biting, sarcastic and having an often aggressively pessimistic worldview (*Rubin et al.*, 2016). It usually denounces

corruption and organizations with contempt and moral outrage (*Thrall and Hibbard*, 1960).

Satire can employ techniques such as exaggeration, the use of vulgarity, humor, creating imaginary societies or fictional universes, absurdity, and alternative versions of history. However, none of these are necessarily satirical in themselves (*Gilmore*, 2017). It suggests progress and the betterment of society, and it advocates that the arts can light the path to improvement (*Colletta*, 2009).

2.1.1 Satire as a figurative language

The identification of figurative language can be a challenging problem not only for computers, but also for human beings (*Onan and Toçoğlu*, 2020). Figurative language differs from literal language in the sense that figurative language expresses its meaning through linguistic nuances such as ambiguity, irony, sarcasm and metaphors, to name a few. Understanding the true meaning of these nuances is dependent on our cognitive abilities, which allow us to reason beyond the syntax of a sentence. As a result, one can imagine that detecting satire automatically is a difficult Natural Language Processing (NLP) task. To determine the true meaning of a text and determine whether it is satire or not, a model must have access to contextual knowledge and be able to rely on various social and cognitive capacities that are difficult to represent computationally (*Frain and Wubben*, 2016).

For example, we show in Example 2.1.1 and Example 2.1.2 two examples of figurative language used in satire. In the first example, Example 2.1.1, one needs some additional knowledge about females allegedly getting grumpy and mad when menstruating to understand the language. In Example 2.1.2, one needs to have knowledge about the saying that all single women have cats. In Norway, it is also a saying that "if I don't get a boyfriend/get married, I will be the single old cat lady". This implies that one needs some background information about the saying to completely make sense of the headline as satire.

Example 2.1.1 –

SISTE: Lanserer tamponger med lykkepille slik at jenter med mensen ikke skal være så sure hele tiden.

JUST IN: Launches tampons with a happiness pill so that girls with menstruation will not be so angry all the time.

Example 2.1.2 –

SISTE NYTT: Evig singel kvinne skaffer seg katt nr 3 og innser at løpet nå er kjørt.

LATEST NEWS: Forever single woman gets cat no. 3 and realizes that she will be single eternally.

Figurative language implies information not grammatically expressed to be able to decode its underlying meaning: if this information is not unveiled, the real meaning is not accomplished and the figurative effect is lost (*Reyes et al.*, 2012). Thus, figurative language processing is one of the greatest challenges in computational linguistics, as the words or expressions possesses a meaning that is different from the literal interpretation (*Reganti et al.*, 2016).

2.1.2 Satire in news

News satire resembles regular news by mimicking the format and style of journalistic reporting. News satire is commonly represented in the Horatian style (*Rubin et al.*, 2016). The stories are generally inspired by real news, and cover the same range of subjects as *i.e.*: politics, sport, and crime (*Rubin et al.*, 2016). Beyond the resemblance, news satire has the intention of reporting news using humor, wit and mockery of people or events. Beyond humor and mockery, satire must also serve a purpose. It is not enough to mock a target; some form of critique or call to action is also required. It is this aspect of censoriousness that separates satire from sheer denunciation (*Rubin et al.*, 2016). News satire is particularly popular on the Web, and specifically in social media in which it is relatively easy to mimic a credible news source, and stories may achieve a wide distribution from almost any site (*del Pilar Salas-Zárate et al.*, 2017).

2.2 Previous work on satire detection

The satire detection problem has attracted some interest in the research community, especially for the English language. However, some have also worked on satire detection in Spanish, Turkish, Romanian, and Italian, to mention some. Regardless, to the best of our knowledge, no research on satire detection has been carried out for the Norwegian language.

2.2.1 Supervised Machine Learning approaches

The following section will give an insight into previous research done on satire detection using supervised machine learning. Supervised machine learning is used to predict a certain outcome from a given input, and where there are examples of labeled input/output pairs. By building a machine learning model from these input/output pairs,

the goal is to make accurate predictions for new, never seen before data. Supervised learning often requires human effort to build labeled datasets, but afterwards automates and often speeds up an otherwise laborious or infeasible task (*Müller and Guido*, 2017).

Burfoot and Baldwin (2009) attempted to determine whether or not news articles can be automatically classified as satirical. The method relied on lexical and semantic features, for instance headlines, profanity, or slang, as well as support vector machines on simple Bag-of-Words features which were combined with feature weighting. They show that an important aspect that helps in detecting satire is the headlines, and that the vast majority of the satire documents were immediately recognisable by only reading the headlines alone. This suggests that their classifier may get something out of having the headline contents explicitly identified in the feature vector. Their best overall F1-score achieved was 79.8%. *Burfoot and Baldwin* (2009) also found that combining SVMs with Bi-Normal Separation (BNS) feature scaling achieved high Precision and lower Recall, and that the inclusion of the notion of "validity" produce the best overall F-score, where validity is the relative frequency of the particular combination of key participants reported in the story.

Frain and Wubben (2016) builds on the work by *Burfoot and Baldwin* (2009). They tested three types of models: A Bag-of-Words (BOW) model using unigrams or bi-grams. A model based on 8 textual features, and a model that combines the BOW model with the 8 textual features. To test their models, they used three classifiers, namely Naïve Bayes (NB), Decision Trees (DT), and Support Vector Classifiers (SVC). Their overall best F1-score predicted on all articles was achieved using the BOW model and the combination of the BOW model and the 8 textual features for unigrams. Both reached a score of 93% with SVC's. They also tested their models on different genres within news, and found that satire detection of political articles obtained highest F1-score with 95% using both BOW and textual features for unigrams in combination with SVC's.

Instead of attempting to determine whether or not newswire articles can be automatically classified as satirical, as approached by *Burfoot and Baldwin* (2009), *Rubin et al.* (2016) went for a more fine-grained feature-representation in combination with machine learning. They propose a SVM based algorithm with 5 predictive features (absurdity, humor, grammar, negative affect, and punctuation). This combination deals with the content directly by detecting language patterns, topicality, sentiment, rhetorical devices, and word occurrences which are common to satire and irony. In their approach, they describe news articles as sparse feature vectors using a topic-based classification methodology with the term frequency and inverse document frequency (TF*IDF) weighting scheme. Like *Burfoot and Baldwin* (2009), *Rubin et al.* (2016)

also found that headlines were essential for detecting satire. Additionally, they found that the final line of each article is relevant to satire detection. This is because the final line often is a "punchline" that highlights absurdities in the story or introduces a new element to the joke. They also found the structure of the syntax (sentence length and complexity) was noticeably different in the satirical and legitimate articles (quotations especially). Their best model achieved a F1-score of 87.0%.

A study by *Toçoğlu and Onan* (2019) on detecting satire in Turkish yielded a F1-score of 89.0%, achieved with the use of support vector machines in conjunction with unigram and term-frequency based representation. They present an empirical analysis on nine different text representation schemes (unigram, bigram, trigram) and term-presence, term-frequency, and TF-IDF weighting schemes, and their combinations. For their empirical analysis, they utilized Naïve Bayes, SVM, logistic regression, and C4.5.

Another study by *Onan and Toçoğlu* (2020) on Turkish news articles presents a machine learning-based approach for satirical text identification. They used LIWC (Linguistic Inquiry and Word Count) in combination with 5 supervised machine learning methods (Naïve Bayes, logistic regression, SVM, Random Forest, k-nearest neighbor). With the use of ensembles feature subsets and the random subspace ensemble of the random forest algorithm, they achieved an accuracy score of 96.92%.

Barbieri et al. (2015a) present a system for automatically detecting satire in Spanish news on Twitter. Their system classifies tweets by relying on linguistically motivated features that aim at capturing not the content but the style of the message. Applying seven classes: frequency, ambiguity, Part-of-Speech, synonyms, sentiments, characters, and slang, in combination with SVM. They show with cross-account experiments (experiments that never share tweets of the same Twitter accounts among training and test sets) that their system detects satire with a good F1-score (best, 81.4%), greatly improving performance with respect to a Bag-of-Words baseline.

Barbieri et al. (2015b) propose an approach for detecting news satire on Twitter in different languages (English, Spanish and Italian). Differently from *Burfoot and Baldwin* (2009), their approach avoids the use of word-based features (Bag-Of-Words). They rely only on language independent features, that are referred to as intrinsic word features since they aim to detect inner characteristics of the words. For their machine learning approach, they applied SVM. Their system was able to recognise if a tweet advertises a non-satirical or satirical news, outperforming a word-based baseline. Moreover they tested the system with cross-language experiments, obtaining interesting results, with an average F1-score of 76.7%.

A study by *del Pilar Salas-Zárate et al.* (2017), used a psycholinguistic-based approach to automatically detect satire on Twitter in Mexico and Spain. They performed

supervised machine learning using Sequential minimal optimization (SMO), Bayes Network learning algorithm, and the C4.5 decision tree. Achieving encouraging results with a F1-score of 85.5% for the Mexican news (using SMO), and a F1-score of 84.0% for the Spanish news (also using SMO). Furthermore, their results confirm the usefulness of adopting the linguistic process, the psychological process, and punctuation marks.

Lastly for the machine learning methods, *Reganti et al.* (2016) presents an approach for automatically detecting satire for various sources, namely Amazon product reviews, newswire documents, and Twitter posts. To build their models, they used 7 sets of features in combination with 5 different classifiers; logistic regression, Random Forest, Support Vector Machine, Decision Tree, and an ensemble of classifiers for better performance. The best model for all three corpora is the ensemble of classifiers in combination with all 7 features. The best F1-score for the product reviews was obtained using all features, and resulted in a score of 77.96%. For the Twitter posts, the best F1-score was also obtained using all 7 features, resulting in a score of 78.16%. For the newswire articles corpus, the best F1-score was 79.02%, also obtained using all 7 features.

2.2.2 Deep-Learning architectures

This section provides background information and theoretical topics related to satire detection using deep learning. Deep learning is considered a branch of machine learning, which is the process of learning not just the relationship between two or more variables, but also the knowledge that governs the relationship and the knowledge that gives the relationship meaning (*Zhang et al.*, 2018).

Yang et al. (2017) proposed a 4-level hierarchical network and utilized attention mechanisms to understand satire at both paragraph level and document level. At the paragraph level, they found that psycholinguistic features, writing stylistic features, and structural features was beneficial. Although satirical news are shorter than true news at the document level, they found that satirical news generally contain paragraphs which are more complex than true news at the paragraph level. The analysis of individual features reveals that the writing of satirical news tends to be emotional and imaginative. Their best overall F1-score was 91.46%, using a 4-Level Hierarchical Network with both Paragraph-level and Document-level linguistic features.

A further development on the dataset presented by *Yang et al.* (2017) is provided by *De Sarkar et al.* (2018). Their approach was to create a model based on two modules, S and D, to detect if an article was satire or not. The S-module creates sentence embedding, taking a sequence of word embeddings as inputs. And the D-module creates a document embedding, which acts as a summarization of the document, taking

sentence embedding as inputs. *De Sarkar et al.* (2018) evaluated their model based on four baselines; SVM word, in addition with char n-grams, unigram and bigrams TF-IDF (*Rubin et al.*, 2016), method learning distributed representation for documents (*Le and Mikolov*, 2014), and a 4-Level Hierarchical Network (*Yang et al.*, 2017). They found that their best model outperforms the baseline models on the dataset. Observing that adding word level syntax information improves the performance only by a small margin. Thus, *De Sarkar et al.* (2018) concluded that at the word level, semantic information is more relevant to capture satire than syntax information. A final conclusion of their two modules (S and D), is that their approach achieves comparable results with already existing research on the topic, without the use of linguistic features reflecting satire. Their best overall F1-score was 91.59% using Convolutional Neural Networks (CNN) in combination with GloVe embeddings and syntactic information.

Following the work of *De Sarkar et al.* (2018), *McHardy et al.* (2019) propose a model based on word embeddings. Different from *De Sarkar et al.* (2018), *McHardy et al.* (2019) model is not hierarchical, and introduces less parameters. Instead, they apply attention to words, rather than sentences or paragraphs. Their corpus consists of almost 330k (320219 regular news, and 9643 satirical) German articles. They pre-trained word embeddings of 300 dimensions on the whole corpus using Word2Vec, which is an algorithm using a neural network model to learn word associations from a large collection of text. The satire detector then provides the feature extractor's representation to a softmax layer that performs a binary classification task. Their model obtains a F1-score of 66.5%. They argue that the majority baseline fails since the corpus contains more regular news than satirical news articles.

Goldwasser and Zhang (2016) researched the satire detection text classification problem using a COMSENSE system, a system that makes predictions by making common-sense interpretations over a simplified narrative representation. Their model was designed to capture behavioral expectations using (weighted) rules, instead of relying on lexical features as is often the case in text categorization tasks (like *Burfoot and Baldwin* (2009)). The COMSENSE system first constructs a graph-based representation of the narrative, denoted Narrative Representation Graph (NRG), capturing its participants, their actions and utterance. Based on the NRG their model makes a set of inferences, mapping the NRG vertices to general categories abstracting over the specific NRG. These abstractions are formulated as latent variables in their model. Testing their model on the corpus of *Burfoot and Baldwin* (2009) and on a corpus retrieved by them. Their best F1-score was achieved on the corpus from *Burfoot and Baldwin* (2009), using their COMSENSE$_Q$ system (using only the entity+quoted based patterns), which resulted in the score of 80.8%.

In addition to a machine learning approach mentioned earlier, *Onan and Toçoğlu* (2020) also evaluated satire identification with the use of deep-learning architectures. They applied 5 deep-learning architectures (Convolutional Neural Network, Recurrent Neural Network, Long Short-Term Memory, Gated Recurrent Unit, and Recurrent Neural Network with attention mechanisms) on three word-embeddings schemes (GloVe, FastText, and Word2Vec). Obtaining a classification accuracy of 97.72% with the recurrent neural network architecture with attention mechanism with the use of the GloVe-based word embedding scheme.

Rogoz et al. (2021) researched the detection of satire for Romanian news articles. They presented two methods. The first one was based on low-level features learned by character-level Convolutional Neural Networks (Char-CNN), and their second method employs high-level semantic features learned by the Romanian version of BERT (Bidirectional Encoder Representations from Transformers). On their test set, they achieved a 73% accuracy and a 71% F1-score with the use of their RoBERT (Romaian BERT model). For their first method, Char-CNN, the model performed a 69.66% accuracy score and a 71.09% F1-score.

Ionescu and Chifu (2021) conducted a study on French news articles to classify if they were satirical or regular news. They propose two cross-domain satire detection tasks, one considering full news articles and another considering only news headlines. As a baseline, they employed two classification methods. One based on low-level features, namely a Presence Bits String Kernel (PBSK) using character n-grams as features, and one-based on a sate-of-the-art language model for French (CamemBERT). On their test set, using full news articles, they achieved a top accuracy rate of 97.48% with CamemBERT and unsupervised domain adaptation (DA). They also found that satire detection on news headlines alone is significantly more challenging. Their best overall accuracy score on headlines being 74.07% using PBSK + DA.

Saadany et al. (2020) studies the detection of satire in Arabic satirical and fake news. Their study experimented with different classification techniques and state-of-the-art deep learning models. Their approach achieved good performance employing Convolutional Neural Networks (CNN). The input layer for the CNN was pretrained FastText word-embeddings trained on Arabic Wikipedia, resulting in a F1-score of 98.49%.

A study by *Apolinario-Arzube et al.* (2020) compares deep-learning architectures and machine learning approaches for satire identification in Spanish tweets. They tested their approaches on two types of input, European Spanish and Mexican Spanish. For the machine learning part they used term-counting features (Bag-of-Words (BoW)) with traditional machine learning (random forest (RF), support vector ma-

chines (SVM), logistic regression (LR), and multinomial Naïve Bayes (MNB)). For the machine learning aspect of the study, they tested the model first on different combinations of word n-grams, and second on different combinations of character n-grams. For the machine learning approaches, the highest accuracy score achieved on the European Spanish tweets was 83.523% using char n-grams with SVM. Whereas, the best accuracy score for the Mexican Spanish tweet was 91.431% using char n-grams and SVM.

For the deep-learning methods, they used word embeddings such as, Word2Vec, GloVe, and FastText in combination with Multilayer Perceptron (MLP), Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs). For both the European Spanish, and the Mexican Spanish, their best model was the combination of FastText and BiGru (which is an improved version of RNNs). The accuracy score for the European Spanish was 81.554%, and 90.524% for Mexican Spanish. They also examined the accuracy score using both European Spanish tweets and Mexican Spanish tweets as input. Their best accuracy score obtained for the full dataset was with char n-grams and SVM, resulting in a score of 85.838%. For the deep-learning methods, FastText and BiGRU also perform best on the combination of both datasets with an accuracy score of 85.429%.

Another deep learning method was done by *Li et al.* (2020) on using both textual and visual cues. Their approach relied on a Vision and Language BERT (ViLBERT), a multi-modal model that processes images and texts in two separate streams (stream consisting of transformer blocks based on BERT and co-attentive layers). They fine-tuned the ViLBERT on a satire detection dataset by passing the element-wise product of the final image and text representations into a learned classification layer. Their model achieved an accuracy score of 93.80% and a F1-score of 92.16%. They argue that ViLBERT performs well because it uses early, deep fusion and has undergone multi-modal pre-training rather than only separate uni-modal visual and text pre-traning.

Lastly, for the deep learning architectures, *Casalino et al.* (2021) presents an approach relying on a deep learning model that tackles the satire detection problem by examining lexical, syntactical, and auxiliary features. They exploited an effective pre-trained embedding tool based on FastText. They tested their corpus on Bidirectional Long Short Term Memory (BiLSTM), Soft Attention Mechanism, Convolutional NNs, and Fully Connected NNs. Achieving their best F1-score using sAttBLSTMConvNet, their main model consisting of five layers; Input Layer, BiLSTM Layer, Soft Attention Layer, Convolutional Layer, and Output Layer. The F1-score obtained was 98.9%.

2.3 Machine Learning VS Deep-Learning

The research presented above shows intriguing results when it comes to the satire detection classification problem, using both machine learning approaches and deep-learning architectures. Studies such as *Onan and Toçoğlu* (2020) and *Apolinario-Arzube et al.* (2020) investigate the satire detection problem using both machine learning and deep-learning. Finding that both methods produce similar outcomes.

The results obtained are also produced in multiple different languages (English, Spanish, Italian, Mexican, German, Turkish, Arabic, Romanian, and French). However, most research has been done on the English language. The best result for English is acquired by *Frain and Wubben* (2016), basing their work on three types of models (1. BOW model, 2. 8 textual features, and 3. Combination of BOW model and the 8 textual features), in combination with Naïve Bayes (NB), Decision Trees (DT) and Support Vector Classifiers (SVC). They obtained a F1-score of 93% using both BOW and textual features for unigrams in combination with SVC's. Even though the best English model produced a score of 93%, the best overall was achieved by *Casalino et al.* (2021) on an Italian corpus, reaching a F1-score of 98.9%. Their model relied on deep learning methods with pre-trained FastText embeddings. In addition, *Saadany et al.* (2020) also obtained a high F1-score of 98.49% on Arabic news. Their model employed a Convolutional Neural Networks (CNN) with pre-trained FastText word-embeddings trained on Arabic Wikipedia as input layer. Other studies, like *Yang et al.* (2017), *De Sarkar et al.* (2018), and *Onan and Toçoğlu* (2020) also obtained F1-score above 90%, respectively 91.46%, 91.59% and 97.72%.

2.4 Our approach

Taking previous studies into consideration, we employ a supervised machine learning approach, using simple machine learning methods. We approach the problem of satire detection as a binary classification task (satire: yes or no). Our methods will rely on the two linear models, Support Vector Machines (linear SVMs) and logistic regression, and one generative classifier, namely Multinomial Naïve Bayes (MNB). Similarly to the studies presented above, we will also utilize evaluation methods such as Precision, Recall, F1-score and accuracy.

Chapter 3

Data and Experimental Setup

This chapter presents both the overview of our dataset and our experimental setup. The dataset was supplied by Web64 [1], which is a tech company that gathers data from Norwegian news sites and blogs. Their mission is to use data to better understand the world and how it is connected. Which is applicable for this thesis' aim in classifying Norwegian articles as satirical or as real news. Through two exports, Web64 downloaded Norwegian news articles, and satiric articles from Norsk Rikskringkasting (NRK), Verdens Gang (VG), 5080, Eavisa, Satiriks and Vredens-Gnag. From the collected dataset, 6322 articles were used for the purpose of this thesis. The articles were sorted to represent all of the sources, as well as balancing the dataset to contain 3161 articles of each genre (real news and satire).

We have decided to use SVM because we believe it is fitting for this research's purposes, where the aim is to separate satire and true news. SVM is a classifier using a line, a plane or a hyperplane to separate classes. The hyperplane divides the input data (training data), and determines which class a given object belongs to (satire/non-satire) (*Müller and Guido*, 2017). It allows for the usage of different kernels to solve linear and non-linear classification models (*Apolinario-Arzube et al.*, 2020). We will apply a SVM model using a linear kernel. Similarly, logistic regression is well suited for discovering the link between features, and can be used to classify an observation into one of two classes. Logistic regression is a discriminative classifier; a model that will try to learn to distinguish the classes. For example: All the satire articles contain the word "Trump" and the non-satire articles do not. If that one specific feature nearly separates the classes, the model is satisfied. If you ask the model what it knows about non-satirical articles, it will say that they don't contain the word "Trump" (*Jurafsky and Martin*, 2020).

In addition to the two linear models, we also use a Multinomial Naïve Bayes (MNB)

[1] https://web64.com/

which is a generative classifier mostly used in text data classification. The goal of MNB is to understand the characteristics of an article that is satire versus an article that is non-satire. The model will "generate" a satire article. The model is first presented with a test article. Then, the system asks whether it is the satire model or the non satire model that best matches the article. Lastly, it chooses the preferred model as its label (*Jurafsky and Martin*, 2020).

Naïve Bayes is efficient in the way that it learns parameters by looking at each feature individually and collecting simple per-class statistics from each feature. MNB assumes count data (that is, that each feature represents an integer count of something, like how often a word appears in a sentence). It takes into account the average value of each feature for each class. To make a prediction, a data point is compared to the statistics for each of the classes, and the best matching class is predicted. This leads to a prediction formula that is of the same form as in the linear models (*Müller and Guido*, 2017).

Logistic regression has a number of advantages over MNB. MNB has overly effective conditional independence assumptions. Consider two features which are strongly correlated or two features that are exactly the same. Adding the same feature, *f1*, twice. MNB will treat both copies of *f1* as if they were separate, multiplying them both in, overestimating the evidence. In contrast, logistic regression is much more robust to correlated features; if two features *f1* and *f2* are perfectly correlated, logistic regression will simply assign part of the weight to *w1* and *w2*. Thus, when there are many significant features, logistic regression will determine a more accurate probability than MNB. Logistic regression generally works better on larger documents or datasets and is a common default. Despite the less accurate probabilities, MNB still often makes the correct classification decision. Furthermore, MNB can work extremely well (sometimes even better than logistic regression) on very small datasets or short documents. Additionally, MNB is easy to implement and very fast to train (*Jurafsky and Martin*, 2020).

The rest of this chapter is structured as followed. Section 3.1 will give an overview of the different news sources. Then, section 3.2.2 present the procedure of splitting the dataset into train, validation, and test sets, as well as a summary of corpus statistics, to give insight into the content of the data. We thereafter in section 3.3, describe which methods were used in this thesis. It will start off with the description on how the data was gathered by Web64. Then, it will move over into how the data process has transpired. Then section 3.4 present how the machine learning models have been applied. The evaluation methods used in this thesis is presented in section 3.5. And lastly, section 3.6 gives an introduction to the different types of input we will run the ML models

on.

3.1 News sources

We define *real* news stories as stories that are known to be true and from well trusted news sources. The *satire* news stories are stories that are from news sources that explicitly state they are satirical and do not intentionally spread misinformation (*Horne and Adali*, 2017).

The real news in this thesis are collected from two of Norway's largest news sources, *Norsk Rikskringkasting* (NRK[2]) and *Verdens Gang* (VG[3]). NRK is the state-owned Norwegian public broadcaster and Norway's largest media company. NRK states that they shall strengthen democracy, the Norwegian language and culture, strive for diversity and quality, and be generally accessible. NRK also writes their news in both Norwegian written forms[4]: Bokmål and Nynorsk. VG is a daily Norwegian newspaper, and is Norway's largest newspaper in terms of number of readers. VG[5] is a news leader in the political field.

The satirical news is collected from four websites that explicitly state they are satirical and do not intentionally spread misinformation: *Vredens Gnag*[6], *Satiriks* [7], *5080*[8], and *Eavisa*[9]. We use these sources as gold labels for satire, as we believe that they guarantee correct labeling as satire. Vredens-Gnag is VG's editorial team for satire and humor. Satiriks is NRK's satire site. 5080 is behind 5080.no and 5080 Nyhetskanalen. 5080 Nyhetskanalen is also the main program on Satiriks which shows a sort of overlap between these two sources. The last one, Eavisa, is an entertainment site focusing on satire and humor. It is only meant for entertainment as well as to put satirical and ironic angles on societal critical issues.

3.2 Dataset

We were provided with two data exports from Web64. The first data export was on May 24^{th} 2021 and the last one on February 21^{st} 2022. In total, there are 6322 valid articles evenly distributed between satire and non-satire articles. For an article to be considered

[2]https://www.nrk.no/
[3]https://www.vg.no/
[4]Gunn Enli, Trond Smith-Meyer, Trine Syvertsen; NRK i Store norske leksikon https://snl.no/NRK
[5]Martin Eide; Norsk presses historie: VG i Store norske leksikon https://snl.no/VG
[6]https://www.vredens-gnag.no/
[7]https://www.nrk.no/satiriks/
[8]https://www.5080.no/
[9]https://eavisa.com/

Source	Number of articles
NRK	1629
VG	1532
5080	1501
Eavisa	1156
Vredens-Gnag	469
Satiriks	35

Table 3.1: Total number of articles from each news source.

valid, the length of the title and text must be greater than 1. This is to ensure no empty documents. In addition, duplicates have been removed.

Table 3.1 shows the distribution of the number of articles from the different sources. The source with the most true news is NRK with 1629 articles and the source with the most satire news is 5080 with 1501 articles. The articles, both satire and true news, are based on current affairs.

3.2.1 Overview of the content

To further reflect on the structure of the news articles and the satire articles, some simple textual features like word count and number of sentences were examined. This was done to see if there is a difference between satire and true news. We utilized Stanza, an open-source tool for advanced natural language processing (NLP) which supports the Norwegian form, Bokmål, to investigate the dataset (*Qi et al.*, 2020).

Looking at the number of words and sentences for the satire and real news documents. The satire articles have a total number of 893700 words, and the real news articles a total number of 1064447 words. As for sentences, satire articles consist of 48586 sentences in total, and real news consist of 56479 sentences. From these numbers, one can see that the real news articles consist of more words and sentences than the satire articles. Which corroborates with findings for the English language by *Yang et al.* (2017), as discussed in Chapter 2.

Examining Table 3.2, one can see that the min and max lengths of the titles are almost identical. However, looking at the min and max lengths of the texts, the satire texts are considerably shorter than the non-satire texts. This again substantiates findings by *Yang et al.* (2017), who also found that satirical news is shorter in length than true news.

As one can view from Table 3.2, the title length is considerably lower than the text length. However, as shown in the literature (see Chapter 2), the title is a very important element of an article for the task of satire detection. The primary aim of a title is to

	Title	Text	Satire title	Satire text	Non-satire title	Non-satire text
Min	8	14	8	14	9	46
Max	146	29471	146	19661	145	29471

Table 3.2: Minimum and maximum size of title and text in terms of tokens.

	Non-satire	Satire	Total
Train	2244	2181	4425
Test	616	649	1265
Validation	301	331	632

Table 3.3: Number of documents in train, test and validation splits.

draw a reader's attention to an article and to capture information in a short glimpse, thus leading to its initial selection or rejection (*Yitzhaki*, 2002).

3.2.2 Train, Validation, and Test sets

To test the model, the data is divided into train, test and validation subsets to prevent overfitting and assess the model more effectively. The train, test, and validation method is a technique to evaluate the performance of a machine learning model. To divide the model into training, test and validation set, sickit-learn's `train_test_split` function was applied (*Pedregosa et al.*, 2011). The training set builds the model, the validation set is used to select the parameters of the model, and the test set is the basis of evaluating the performance of the selected parameters (*Müller and Guido*, 2017).

The training set consists of 70% of the dataset, the validation set of 10%, and the test set of 20%. In Table 3.3 the number of articles in the train set, test set, and validation set is presented. From the table, one can see that the satire and non-satire articles are almost evenly distributed in the train, validation, and test sets.

In Table 3.4 the distribution of the distinct sources in the training set, test set, and validation set is shown. In the train set, NRK and VG are represented with over 1000 articles each. For the satire news, 5080 and Eavisa are most represented, with 1037 articles from 5080 and 808 articles from Eavisa. Satiriks, with only 22 articles, is the least represented source in the train set. However, as previously stated, the majority of the content on Satiriks comes from 5080.

The number of articles is closely distributed the same way as the training, testing, and validation set is. With approximately 70% of the articles from each source in the train set, and roughly around 20% test set and 10% for the validation set.

	5080	Eavisa	Satiriks	Vredens-Gnag	VG	NRK	Total
Train	1037	808	22	314	1099	1145	4425
Test	297	236	5	111	300	316	1265
Validation	167	112	8	44	133	168	632

Table 3.4: Distribution of the different sources in train, test and validation

3.3 Data preprocessing

Data preprocessing is a major step in the data mining process. The process of gathering information and making it useful, filtrating away noise and empty values. In other words, data preprocessing is transforming data into a form that computers can easily work with (*García et al.*, 2015).

To obtain information, Web64 has developed a media monitoring platform that gathers Norwegian news articles published online. The news is retrieved through multiple different channels. They have built a database consisting of thousands of RSS-feeds, which makes it possible for a browser to gather news consecutively from the Internet. In addition, Web64 has a crawler that continuously checks Norwegian websites and news sites for new links to collect. Using this method, Web64 has contributed with the dataset applied in this thesis.

After collection, the data was converted into a CSV format. The web page menus, headers and footers were not incorporated. The CSV file consists of 5 columns and 6322 rows. Each row corresponds to one article. The columns store the articles ID, title, text, source, and category. The ID is the reference number of each document in the dataset. Title and text are equivalent to the articles' titles and texts. Source stores where the articles are gathered from (NRK, VG, Satiriks, 5080, Eavisa, and Vredens-Gnag), and category labels if an article is satire (s) or non-satire (ns).

Once the CSV file was created, we transformed it into a Pandas DataFrame structure that can be used with Python. The CSV file was constructed to make extracting the desired features easier. Then, the CSV file was divided into training, test, and validation sets using sickit-learns `train_test_split function` (*Pedregosa et al.*, 2011). The training set and the validation set were both converted to TF-IDF feature vectors, by applying `fit` (learn vocabulary and IDF from training set) and `transform` (return document-term matrix). After the corpus was prepared, the models were then trained using three different classification approaches: Linear SVM, Naïve Bayes, and logistic regression.

After the models were trained, we evaluated their performance first on the validation set, then on the test set. We used Precision, Recall, F1-score (both class-level and micro

average) and accuracy as evaluation metrics for both the satire and non-satire articles.

We also generated confusion matrices to investigate how many articles the model correctly or incorrectly classified. The evaluations of our models can be found in Chapter 4. All the results were obtained using scikit-learns `classification report` method and `confusion matrix` method (*Pedregosa et al.*, 2011).

The whole process can be viewed in Figure 3.1 below.

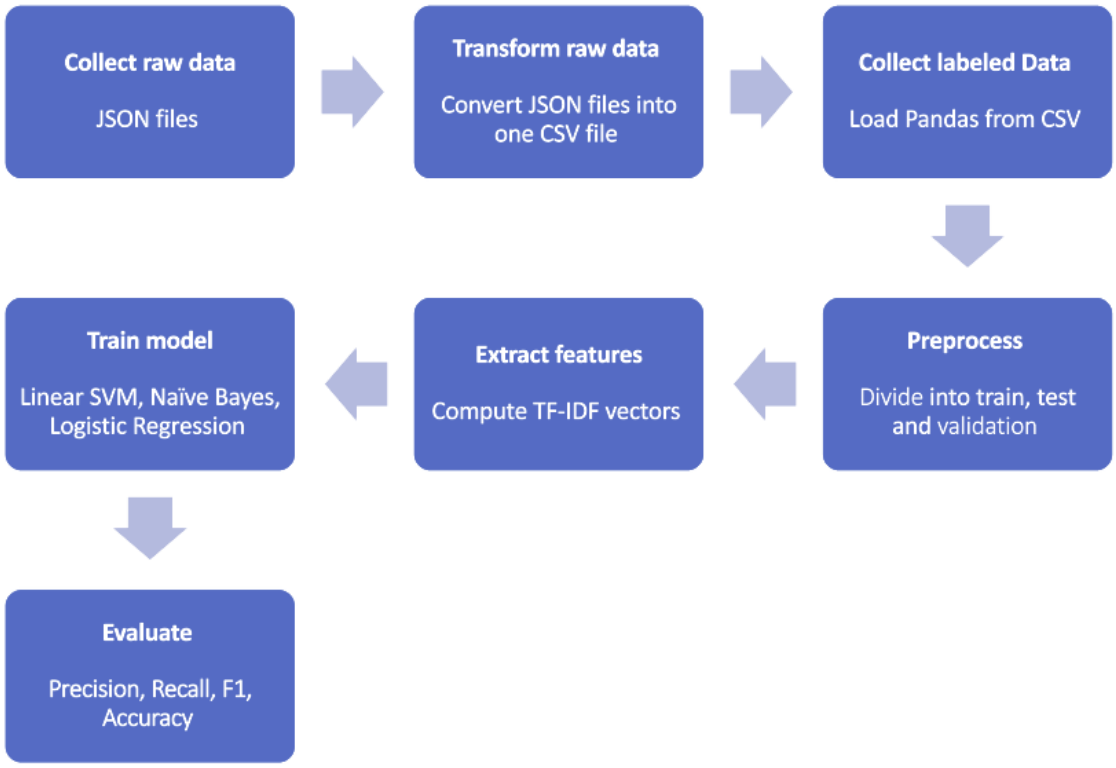

Figure 3.1: News satire detection pipeline for distinguishing satirical from real legitimate news.

3.4 Implementation

Using a binary-based classification methodology with the TF-IDF weighting scheme, the news articles are represented as feature vectors. The term frequency-inverse document frequency (TF*IDF) method re-scales features by how informative they are expected to be. Any term that appears frequently in a single document but not in many others in the corpus is given a high weight by TF-IDF. Because, if a word appears often in a document but not in many others, it is likely to be particularly descriptive of that document's content (*Müller and Guido*, 2017).

For TF-IDF, Scikit-learn's (*Pedregosa et al.*, 2011) `TfidfVectorizer` was applied. The `TfidfVectorizer` takes in text data and extracts bag-of-words features as well as doing the TF-IDF transformation. The `TfidfVectorizer` in this research takes the parameters `min_df` and `encoding`. `Min_df` is a float or an integer. When building

the vocabulary it ignores the terms that have a document frequency strictly lower than the given threshold (the `min_df`). The threshold in this study is set to be 5. That is, removing words that occur in less than 5 documents. After the corpus is fitted to the `TfidfVectorizer`, the binary classifiers predict the result. The results of the news corpus were achieved using a linear SVM classifier (assigning positive instances to satire), Naïve Bayes (understanding the characteristics of an article), and logistic regression (discovering the link between features).

3.5 Evaluation

The metrics used to evaluate the performance of our ML models are Precision, Recall, F1-score, and accuracy, predicted for both the satire news and the true news.

The evaluation methods deployed in the thesis, is the same as for the previous research presented in Chapter 2. In addition, a confusion matrix will also be calculated for all ML models. The dataset contains regular news and satirical news articles. The advantage of using a labeled dataset, is that the number of articles is set, and the distribution of real news and satire news is known. In this thesis, all "relevant elements" will be the satire articles, and the regular news articles will be "not-relevant elements". This will help distinguish between the desired articles and the not desired articles, and make it possible to calculate. In what follows we give a short description of each metric.

Precision: Precision calculates the percentage of the items that the system detected (*i.e.,* the system categorized them as positive) that are in fact positive (*i.e.,* are positive according to their labels, here satire or non-satire) (*Jurafsky and Martin*, 2020).

Recall: Recall calculates the percentage of items actually present in the input that were correctly identified by the system. Recall is therefore a measure of effectiveness in retrieving performance and can be viewed as a measure of effectiveness in including relevant items in the retrieved dataset (*Jurafsky and Martin*, 2020).

F1-score and Accuracy: After predicting Precision and Recall, one can calculate the class-level and macro average F1-score, as well as accuracy score.

Accuracy is the measure of all correctly identified cases. Accuracy is usually not applied for text classification tasks, due to the fact that it does not work well with unbalanced datasets (*Jurafsky and Martin*, 2020). However, in this case it can be applied, since we have a balanced dataset consisting of 3161 satirical articles and 3161 true news articles.

Confusion Matrix			
	Satire	True News	
Satire	True Positive	False Positive	Precision = tp/tp+fp
True news	False Negative	True Negative	
	Recall = tp/tp+fn		Accuracy = tp+tn/tp+fp+tn+fn

Table 3.5: Confusion matrix as presented in and by Jurafsky and Martin (2020).

Confusion Matrix: A confusion matrix for each model will also be calculated, see Table 3.5. A confusion matrix is a table for visualizing how well an algorithm performs with respect to the labels, using two dimensions (system's outputs and labels), and each cell labeling a set of possible outcomes (*Jurafsky and Martin*, 2020). In the satire detection case, true positives (TP) are articles that are indeed satire, which the model correctly said were satire. True negatives (TN) are true news correctly classified by the model as true news. False negatives (FN) are articles that truly are satire but the system have incorrectly labeled as true news. And lastly, false positives (FP) are articles that indeed are true news, but have incorrectly been classified as satire.

3.6 Experiments

To test our second research question (RQ2), that is which types of input achieve the highest classification scores, we focus on three different tasks:

1. Task 1: use only title as input.
2. Task 2: use only text as input.
3. Task 3: use a combination of title and text as input.

Task 1 is running the models using only the title column as input. Task 2 is running the models using only the text column as input, and Task 3 is running the models using both the title and the text column as input. From now on, they will be referred to as Task 1, Task 2 and Task 3 respectively.

In this thesis, the classification models will correspondingly be trained on the title, the text, and title and text combined.

Previous research states that headlines are important in detecting satire (*Burfoot and Baldwin*, 2009; *Rubin et al.*, 2016). In addition *Ionescu and Chifu* (2021) also studied how well satire detection would be for only title as input, resulting in a F1-score of 74.07%. It will therefore be relevant to have our models be carried out on only title, to consider if our models will produce a similar outcome as previous work.

Additionally, it would be interesting to explore how our models will perform on text input only. *Rubin et al.* (2016) found that the structure of the syntax (sentence length and complexity) was noticeably different in satirical and true news articles. Similarly, *Yang et al.* (2017) found that satirical news generally contains paragraphs which are more complex than true news at the paragraph level. This makes it interesting to investigate only the text in the articles further, and whether or not they can be automatically detected as satire.

Furthermore, the models will also be processed on the combination of title and text as input (Task 3). It will then be compared and analyzed with the result achieved from Task 1 – title as input, and Task 2 – text as input, to see how this may affect our models.

Chapter 4

Results and Discussion

In this chapter, we present and analyse the results achieved by our three ML models, Naïve Bayes, SVM, and logistic regression, for our three tasks, Task 1, Task 2, and Task 3. As introduced in chapter 3, our three tasks have three different inputs: (i) Task 1: title as input, (ii) Task 2: text as input, and (iii) Task 3: title and text as input. The models are evaluated using Precision, Recall, F1-score, and accuracy

We explore the most informative features for each model in the test set. This tells us which words were the most important during classification of both satire articles and for true news articles. We also perform an extensive error analysis, looking at the miscclassifications of our models, and analysing why the ML models may have wrongly classified an article text or title, to see how the models may be improved. Additionally, as mentioned in chapter 3, a confusion matrix is constructed for all three Tasks for all three models.

The rest of this chapter is structured as followed. First, in Section 4.1 and Section 4.2, we give an overview of the results obtained by our ML models for all three tasks, as well as a discussion of our results. We start by presenting the results achieved on the validation set, then we present the results on the test set. Section 4.4 presents the error analysis performed on the test set. Looking at why the ML models may have wrongly classified an article text or title, to see how the models may be improved. At the end, in Section 4.12 we present our results compared with previous research. Lastly, we give in Section 4.13, a discussion of the limitations of our work.

4.1 Performance on the validation set

We give an overview and analysis of the performance of our three models, SVM, Naïve Bayes, and logistic regression for the task of classifying our validation dataset into satire or non-satire classes. We are performing the tasks on three types of input: title

	P_S	P_NS	R_S	R_NS	F1-score_S	F1-score_NS	Acc
Naïve Bayes	0.79	**0.74**	**0.75**	0.78	**0.77**	**0.76**	**0.77**
SVM	0.80	0.72	0.71	0.80	0.75	**0.76**	0.75
Logistic Regression	**0.81**	0.71	0.70	**0.82**	0.75	**0.76**	0.76

Table 4.1: Results from Task 1 – title as input. Where P stands for Precision, R for Recall, and Acc for Accuracy.

alone, text alone, and a combination of title and text.

4.1.1 Task 1: title as input

Table 4.1 displays the results achieved from running the three models with title as input. Naïve Bayes achieves the best accuracy score with 77%. But, all three classifiers yielded similar results. There was only a 2% difference in results, with the lowest accuracy score being 75% with the use of SVM.

For each of the three models, the macro average was computed. The findings obtained the score of 77% for Naïve Bayes, 75% for SVM, and 76% for logistic regression.

The best Precision score for satire news was achieved using logistic regression, scoring 81%, and the best Recall score was achieved with Naïve Bayes, scoring 74%. For true news, Naïve Bayes achieved the highest Precision with 74%, and logistic regression achieved the highest Recall with 82%. Table 4.1 displays that the highest F1-score for satirical articles, with a score of 77%, is when applying Naïve Bayes. For the true news, the F1-score for all three ML models reached 76%.

The model that produces the best overall result for Task 1 - title as input, is Naïve Bayes. However, the gap between scores is not large. The difference between the best overall F1-score for satire news and the best overall F1-score for real news is only 1%. Calculating the average F1-score from all three models, then the F1-score for true news (76%) is higher than for satire news (75,6%), regardless of the highest F1-score being produced for satire news. Despite this, the difference is minimal, and the models achieve almost the same score for both satire and true news.

4.1.2 Task 2: text as input

The results of training our models using text as input are shown in Table 4.2. Using text as input significantly improved the performance of the classifiers compared to utilizing only titles as input - with a maximum accuracy score of 97% using SVM. However, looking at all accuracy scores, there is only a 3% difference between the models,

	P_S	P_NS	R_S	R_NS	F1-score_S	F1-score_NS	Acc
Naïve Bayes	0.90	**0.98**	**0.98**	0.88	0.94	0.93	0.94
SVM	**0.97**	0.97	0.97	**0.97**	**0.97**	**0.97**	**0.97**
Logistic Regression	0.94	0.97	0.97	0.93	0.95	0.95	0.95

Table 4.2: Result from Task 2 – text as input. Where P stands for Precision, R for Recall, and Acc for Accuracy.

	P_S	P_NS	R_S	R_NS	F1-score_S	F1-score_NS	Acc
Naïve Bayes	0.90	**0.97**	**0.98**	0.88	0.94	0.92	0.93
SVM	**0.97**	**0.97**	**0.98**	**0.96**	**0.97**	**0.97**	**0.97**
Logistic Regression	0.94	0.96	0.96	0.93	0.95	0.94	0.95

Table 4.3: Result from Task 3 – title and text as input. Where P stands for Precision, R for Recall, and Acc for Accuracy.

whereas Naïve Bayes produced the lowest score of 94%.

For Task 2: title as input, SVM achieves the best Precision for the satire news (97%) and the highest Recall for true news (97%). For Precision for true news, and Recall for satire news, Naïve Bayes produces the best scores with 98% for both.

The top F1-scores for satire news and real news are both attained using SVM, with a score of 97%. The F1-scores for Naïve Bayes and logistic regression are nearly identical for both inputs, with the exception of the F1-score for Naïve Bayes, which is 1% lower for real news (93%) than for satire news (94%). In addition, the macro average was also computed for Task 2. The scores produced was 94% for Naïve Bayes, 97% for SVM, and 95% for logistic regression.

4.1.3 Task 3: title and text as input

The results from combining the two previous types of inputs (title and text), is shown in Table 4.3. Again, the SVM algorithm achieves the highest level of accuracy, scoring 97%. The result achieved with the combination of title and text is closely similar to the result achieved in Task 2 – text as input (as shown in Table 4.2). The only difference in accuracy scores is for the Naïve Bayes classifier, with a 1% lesser accuracy score for title and text (93% vs. 94% from Task 2). It can be seen in both Table 4.1 and Table 4.2, that the score is significantly higher when text is applied as input, than the score is when only using title. Due to this, it is quite likely to believe that the text played a more significant influence in determining the final classification.

Viewing Table 4.3, both Naïve Bayes and SVM produced good Precision for true news (both 97%) and Recall for satire news (both 98%). Looking at Precision for satire

and Recall for true news, SVM achieved the best result with 97% for Precision and 96% for Recall. Overall, SVM produced the highest performance, achieving a 97% F1-score for both satirical and real news. Naïve Bayes yielded the lowest result with a 94% F1-score for satire and 92% for true news. For the macro average the score is 93% for Naïve Bayes, 97% for SVM, and 95% for logistic regression.

4.1.4 Discussion

For Task 1 the Naïve Bayes model accomplished the best result. For Task 2 and Task 3 the best performing model is SVM for both tasks.

Burfoot and Baldwin (2009) and *Rubin et al.* (2016) both found that headlines were important when detecting satire, and that most articles could be classified based on title alone. Results from the ML models used in this study indicate that longer text lengths achieves better results. This substantiates findings by *Ionescu and Chifu* (2021) who also got an accuracy score of above 90% for text as input, and an accuracy score of under 75% for title as input.

The average title accuracy score was 76%, while the average text accuracy score was 95.36%. The average score is 95% when combining both the title and the text. Therefore, it is reasonable to assume that title offers little to no value to the classifier when the two are joined.

Logistic regression provides the best Precision for the satire news in Task 1. For Tasks 2 and Task 3 SVM provides the highest Precision for satire news. For all three tasks, Naïve Bayes produced the highest score for satirical news articles on Recall. SVM achieves the highest F1-scores for satire news for Tasks 2 and 3.

Considering the true news, the Naïve Bayes algorithm achieves the highest Precision scores for all three tasks. For Recall, SVM is best on Task 2 and 3, whereas logistic regression is best for Task 1. SVM yields the highest F1-score for true news.

Overall, Naïve Bayes produced best scores for Task 1, while for Task2 and Task 3, SVM is best.

4.2 Performance on the test set

For the purpose of classifying our test dataset into satire and non-satire classes, we present an overview and analysis of the performance of our three models Naïve Bayes, SVM, and logistic regression. The task is being done on three different sorts of input: only title as input, only text as input, and a combination of titles and texts as input.

	P_S	P_NS	R_S	R_NS	F1-score_S	F1-score_NS	Acc
Naïve Bayes	0.76	**0.75**	**0.76**	0.74	**0.76**	0.74	0.75
SVM	**0.79**	0.73	0.72	**0.79**	0.75	**0.76**	**0.76**
Logistic Regression	0.78	0.73	0.72	**0.79**	0.75	**0.76**	0.75

Table 4.4: Results on the test set, Task 1 – title as input. Where P stands for Precision, R for Recall, and Acc for Accuracy.

	P_S	P_NS	R_S	R_NS	F1-score_S	F1-score_NS	Acc
Naïve Bayes	0.91	0.98	0.98	0.90	0.94	0.94	0.94
SVM	**0.98**	**0.99**	**0.99**	**0.98**	**0.98**	**0.98**	**0.98**
Logistic Regression	0.95	0.97	0.97	0.95	0.96	0.96	0.96

Table 4.5: Results on the test set in Task 2 – text as input. Where P stands for Precision, R for Recall, and Acc for Accuracy.

4.2.1 Task 1: title as input

The result on Task 1 is presented in Table 4.4 below. SVM produced the highest accuracy score, achieving a score of 76%. However, looking at the scores given by Naïve Bayes and logistic regression they performed only 1% less accurately than SVM, with an accuracy score of 75%.

For satire news, SVM achieved the highest Precision score of 79%, while Naïve Bayes achieved the highest Recall score of 76%. For the true news, Naïve Bayes achieved the best result for Precision with 75%. Both SVM and logistic regression yielded the highest Recall score for true news (79%). The highest F1 score for both satire and true news was 76%. However, Naïve Bayes got the highest F1-score for satire, whereas SVM and logistic regression had the highest F1-score for real news. Examining the F1-score for both satire and true news reveals that the average score attained by all three models is 75.3%. Macro average F1 was likewise predicted, yielding the scores of 75% for Naïve Bayes, 76% for SVM, and 75% for logistic regression.

4.2.2 Task 2: text as input

Executing the models on Task 2 achieved an accuracy score of 98% with the use of SVM, visualized in Table 4.5. Naïve Bayes produced the lowest accuracy score with 94%, and logistic regression the next best accuracy score with a score of 96%.

Looking at Table 4.5 for Task 2 the highest scores were all generated using SVM. Precision for true news and Recall for satire news both received a score of 99%, the highest score produced. Both the satire news and the true news, achieved a F1-score of 98% with SVM. In comparison Naïve Bayes scored a 91% Precision score for satire and

	P_S	P_NS	R_S	R_NS	F1-score_S	F1-score_NS	Acc
Naïve Bayes	0.91	**0.98**	**0.98**	0.89	0.94	0.93	0.94
SVM	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**
Logistic Regression	0.95	0.96	0.96	0.94	0.96	0.95	0.95

Table 4.6: Results on the test set for Task 3 – title and text as input. Where P stands for Precision, R for Recall, and Acc for Accuracy.

a 90% Recall score for true news, which is lower than for SVM and logistic regression. In addition, the macro average F1 was computed. Achieving the score of 94% for Naïve Bayes, 98% for SVM, and 96% for logistic regression.

4.2.3 Task 3: title and text as input

Lastly, the models were trained using the input representations of Task 3, resulting in an accuracy score of 98% with the use of SVM. Logistic regression accomplished an accuracy score of 95%, and Naïve Bayes reached an accuracy score of 94%.

Similarly to Task 2, SVM achieved the best scores on Task 3. Viewing Table 4.6, Naïve Bayes also produces some of the highest scores, but SVM is overall best. The scores achieved by SVM is identical for all calculations, with a score of 98% for Precision, Recall, F1-score, and accuracy for both satire news and true news. In addition, Naïve Bayes also produced a 98% Precision score for true news and Recall for satirical news. Nevertheless, on all the other scores, Naïve Bayes performs the lowest. Complementary to Task 1 and Task 2, the macro average F1 produced the same result as the accuracy score for all three models, with 94% for Naïve Bayes, 98% for SVM, and 95% for logistic regression.

4.2.4 Discussion

For the test set, SVM performed the best overall accuracy score for all three tasks. In addition, SVM performs best on Precision regarding satirical articles. For the true news articles, Naïve Bayes and SVM both perform equally good for Precision. Highest Recall score for satire news is achieved by both Naïve Bayes and SVM, and for the true news SVM is best. The best F1-score for both Task 2 and Task 3, is achieved by SVM for both satire news and true news. For Task 1, Naïve Bayes achieves the highest score for satirical news, and SVM and logistic regression both perform equally well for true news.

According to *Ionescu and Chifu* (2021) detecting satire solely based on headlines is a difficult task. Achieving comparable outcomes to our models. The accuracy of their

models for detecting satire through the use of only headlines is less than 75%, with the best being 74.07%. Our models produce similar, yet, slightly better results, with the best score on the test set for Task 1 being 76% using SVM.

Similarly to our findings, *Ionescu and Chifu* (2021) also achieved accuracy scores of over 90%, taking only text as input. Their best score achieved on the test set was with the use of CamemBERT (pre-trained language model for French) in conjunction with DA (unsupervised domain adaptation), reaching a score of 97.48%. Our overall best score was achieved on Task 2 and Task 3 on the test set, using simple machine learning models, where SVM produced the best outcome. This indicates that it is possible to achieve good results with simple ML models and merely the text of an article.

4.3 Summary of findings

We summarize here our main findinds, focusing on the test set. From our experiments, it is clear that the SVM model provided the best performance for all taks. Examining all task results obtained by the validation set and the test set, it is clear that there is little variation in the resulting scores (only 1% - 3% difference in scores). Only Task 1 has a higher score on the validation set than it did on the test set. For Tasks 2 and 3, the test set had greater scores than the validation set.

When running the ML models on Task 1 on the test set, the result of 76% is achieved with SVM, which is 1% lower than running the models on the validation set (77%). In addition, the models yielding the best result are different for the two sets. For Task 1 on the validation set, Naïve Bayes achieved the best score, however, for Task 1 on the test set, SVM had the overall best score. The score achieved with Naïve Bayes on the test set was also 2% lower than achieved on the validation set. The result for logistic regression was similarly poorer on the test set than on the validation set, 75% reached on the test set and respectively 76% for the validation set. All things considered, the ML models achieve worse on the test set than for the validation set. Regardless, the performance is nearly identical.

In contrast to Task 1 – title as input, Task 2 – text as input performed better on the test set compared to the validation set. With increasing the overall best score with 1%, from 97% to 98%. Both results were obtained with SVM, which indicates that SVM is the best model to use for detecting satire with longer texts. As for the other ML models, Naïve Bayes produced the same accuracy with 94%, while logistic regression increased the accuracy score with 1%, from 95% to 96%.

Similarly to Task 2 – text as input, Task 3 – title and text as input, also generated better scores on the test set compared to the validation set. With an increase from 97%

on the validation set to 98% on the test set. Furthermore, Task 2 and Task 3 on the test set produced similar accuracy scores. The only difference can be seen in logistic regression, with 1% lower score on Task 3 than for Task 2. Comparing Task 3 on the validation set to Task 3 on the test set, reveals a 1% improvement for all models for the test set, except for logistic regression, which score stayed unchanged.

When looking at both the validation set and the test set, the overall best model is SVM. SVM is the model that has achieved the highest scores for Precision, Recall, F1-score, and accuracy for both satirical and real news. Next best model is Naïve Bayes. And the model that has produced the lowest performance is logistic regression. However, it is worth mentioning that all three ML have yielded comparable findings.

Considering the Precision scores for the validation set and test set for both satire and regular news, the difference within each task is similar. The biggest difference in Precision score for satirical and true news can be viewed in Task 1 for the validation set. Where the best Precision for satire is 81%, whereas the best Precision for true news is 74%.

Furthermore, the best scores for Recall are also similar within all tasks for both satire and regular news. Again, the biggest difference can be seen in Task 1 for the validation set. Where Recall for satire is 75% and for true news is 82%. Considering the F1-scores within each task the scores are similar for both satire and true news.

4.4 Error Analysis

In order to get a better understanding of what each model has learnt during training, we perform an error analysis of the classification errors made by each model on the test set. This section will present the error analysis completed on the test set. *Rundell et al.* (2022) define *Error Analysis* as:

Definition 1 *"The process or activity of looking at errors in order to find out what they are, why they are happening and what can be done to prevent them."*

Error analysis is the process of analyzing the uncertainty associated with a measurement. Evaluating the work of the obtained data in order to determine where anything went wrong, as well as discussing and determining viable solutions for how the models might be improved in order to make accurate predictions the next time.

We here do different types of analysis. First, we look a the confusion matrices for each model, and explore which classes seems to cause most problem during classification. Second, we look into the top 10 most informative features for each model. Third, and this is solely done for Task 1, we manually annotate the titles that were missclassified by all three models and manually categorize them into topics. This was done to

investigate if there exist topics where it is more difficult to separate between satire and non-satire.

We extensively report our analyses for each model and task, and give in what follows a detailed description of the missclassifications of our models.

4.5 Task 1 – test set

We manually analysed the set of incorrectly classified titles for Task 1 – title as input, for the test set. All three models have a total of 300 misclassifications. We decided to focus on the 196 titles that were wrongly classified by all three models.

We here focus on Task 1 – title as input, both due to time constraints as analysing title is less time-consuming. But also since it is clear that all our models performed poorly on this task compared to the two other tasks.

We performed an extensive manual error analysis for Task 1 on the test set. The analysis was made by three human annotators, including the author of the thesis. All sentences were analysed and categorized by all three annotators simultaneously. Categories were assigned only upon majority agreement, we therefore do not provide any inter-annotator agreement.

The manual annotations have resulted in 22 categories of titles. These represent what seems to be covered by the news, and was done in order to shed light on which topics seems to be more difficult to differentiate between using our simple ML models. Table 4.7 shows the annotated categories, and how many of the satirical and true news of each category were missclassified by our models.

Naïve Bayes has 314 incorrectly classified titles, whereas 155 were predicted as satire, but in reality were true news (FP), and 159 were predicted as true news but in fact were satire (FN). SVM has 309 misclassified titles, whereas 181 were predicted as satire but were true news (FP), and 128 were classified as true news but were satire (FN). Logistic regression has 311 misclassified, whereas 182 were classified as satire but were true news (FP), and 129 were predicted as true news but were satire (FN).

In total, the models have predicted more articles to be satire when they in fact were true news (FP), then it has predicted true news to be satire (FN). For Naïve Bayes the model predicted almost the same amount of articles to be FP (predicted satire, but are true news) and FN (predicted true news, but are satire) (see Figure 4.1 below). For SVM and logistic regression, the models predicted approximately 180 as being satire, but in reality is true news (FP), and around 130 as true news, but in reality is satire (FN) (see Figure 4.3 and 4.4).

For the 196 misclassified titles made by each model, 112 title were labeled true

Categories	True News	Satire
Politics	16	18
Health politics	14	2
Sport	7	17
Agurknytt	11	12
Entertainment	-	6
Health	2	1
Economics	4	4
Foreign policies	-	9
Crime	4	-
Weather	3	5
Chronicle/Opinion	1	1
Celebrity	6	7
Research	2	1
General news	9	2
The Noble Peace Prize	1	1
Humor	-	16
History	-	1
Foreign	-	2
Culture	-	3
Quiz	1	-
Review	3	-

Table 4.7: The different categories manually assigned to the missclassified titles by all three models.

news when they actually were satire (FN), and 84 were predicted to be satire but were in reality true news (FP).

4.5.1 Naïve Bayes

Figure 4.1 presents the confusion matrix for Naïve Bayes for Task 1 on the test set. It shows that the model correctly classified 951 titles (494 as true positives (TP), and 457 as true negatives (TN)), and wrongly classified 314 titles (159 as false negative (FN), 155 as false positive (FP)). Looking at the misclassified headlines, the distribution of the wrongly labeled titles are almost even. With only 4 more FN, than FP. This indicates that the Naïve Bayes model struggles equally at classifying a title as satire or true news.

In Table 4.8 the most informative features for Naïve Bayes with title as input is presented. The words are sorted by their score from top to bottom. Looking at Table 4.8, most satirical features are stop words. A reason for why the most informative features for Naïve Bayes only are stop words may be because Naïve Bayes uses the frequency of each word (*Jurafsky and Martin*, 2020). And since stop words are used multiple times in a text, it is fair to assume that they have been given higher weights. The words *siste* (last/latest) and *nytt* (news) are often used together with each other,

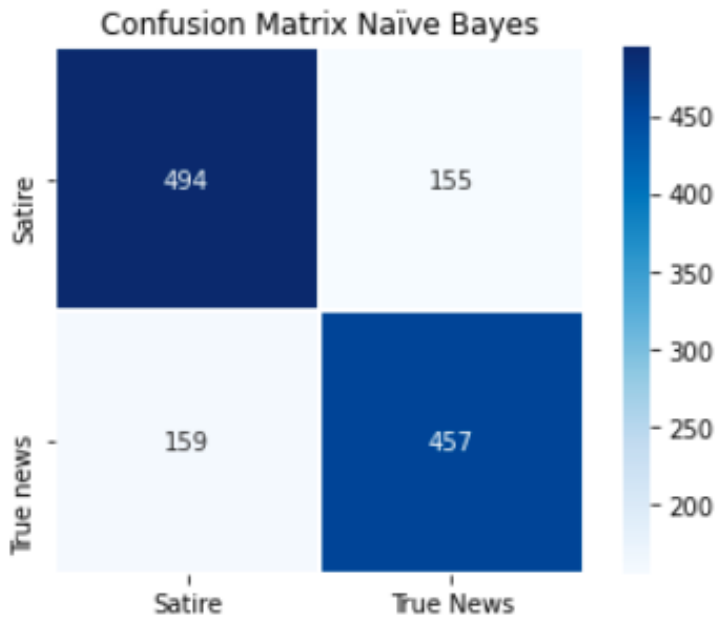

Figure 4.1: Confusion matrix for Task 1 using Naïve Bayes.

Figure 4.2: Example headlines from the satirical news outlet Eavisa.

especially in Eavisa.

Figure 4.2 presents an excerpt from Eavisa and some of their headlines. Two of the headlines, the first and the last, both start with the words *SISTE NYTT* (LATEST NEWS). The word *siste* (latest) occurs in the top ten most informative features for all three models (see Tables 4.8, 4.9, and 4.10). The word *nytt* (news) occur in the top ten most informative features for Naïve Bayes (see Table 4.8) and logistic regression (see Table 4.10). The second headline start with the word *STUDIE* (STUDY), referring to a study stating that couples traveling to Hawaii with private planes are more happy than couples traveling with *Danskebåten* (the Danish boat). Travelling with Danskebåten is viewed as a stereotypical destination for Norwegians. The third headline start with *KVINNE FORTVILER* (WOMAN DESPAIRS) and is about a woman allegedly being kicked out of the friend group because she has talked about christmas since September.

True news		Satirical	
Feature	Translation	Feature	Translation
andersen	andersen	på	on
angrep	attack	for	for
anmeldelse	review	til	to
avlyse	cancel	av	off
berg	berg	siste	last/latest
bodø	bodø	med	with
boris	boris	nytt	news
casper	casper	er	is
djokovic	djokovic	som	as
e6	e6	og	and

Table 4.8: Top 10 most informative features for Task 1 for true news and satire using Naïve Bayes.

Looking at Table 4.8, one can observe that the most informative features for Naïve Bayes for true news seem to cover international and national matters. For international matters, words such as *Boris* and *Djokovic* occurs. For the national matters, words such as *E6*, *Andersen*, *Berg*, and *Bodø*, cover national events. With E6 referring to one of the largest highways in Norway. Andersen and Berg a typical Norwegian names, and Bodø ia a Norwegian city.

When observing the misclassified headlines, some of the ones that were wrongly classified as true news, when in reality it was satire, were headlines starting with CAPS LOCK, followed by a colon ∴. This substantiates findings by *del Pilar Salas-Zárate et al.* (2017) who also found that colons are frequently used to denote satire.

For example, the first title listed below, Example 4.5.1, is from Eavisa. The title has been classified as a *sport* title, and is stating that Norwegian Olympic athletes are selling their asthma medicine on the black market. The second title, Example 4.5.2, is from Vredens-Gnag. The title has been classified as a *humor* title. It describes a virus attack, were the Home Guard has begun shooting at computers. In the last title, Example 4.5.3, which is another example from Eavisa. The title has been classified as *humor*. It explains how an increasing amount of mothers-in-law are locked in cars.

Example 4.5.1 –

NY OL-SKANDALE: Norske OL-utøvere selger astmamedisin på svartebørsen til 9000 euro stk!

NEW OLYMPIC SCANDAL: Norwegian Olympic athletes sell asthma medicine on the black market for 9,000 euros each!

Example 4.5.2 –

VIRUSANGREPET: Heimevernet har begynt å skyte PC-er

VIRUS ATTACK: The Home Guard has started firing at computers

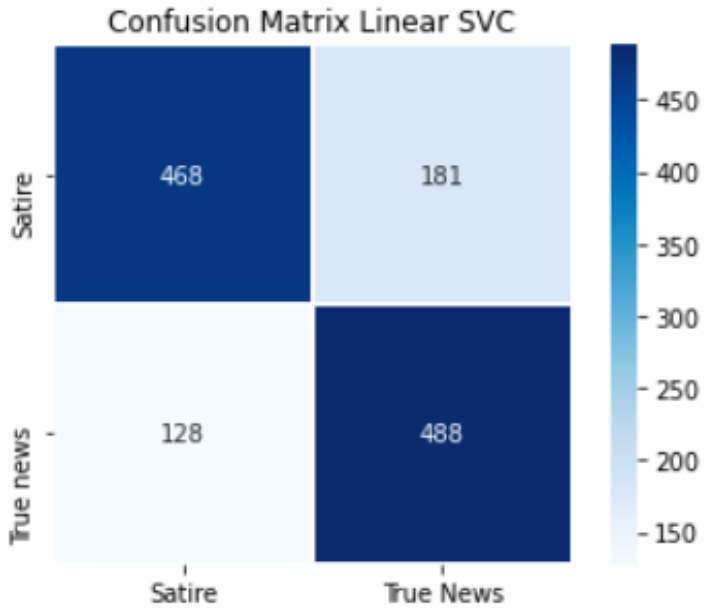

Figure 4.3: Confusion matrix for Task 1 using a linear SVM.

Example 4.5.3 –

POLITIET ADVARER: Stadig flere svigermødre sitter innelåst i glovarme biler.

POLICE WARNING: More and more mothers-in-law are locked in hot cars.

Looking at the headlines for true news, if the headlines start with CAPS LOCK and then a colon *:*, they often start with names or instances where the abbreviation is in all capital letters. Consider Example 4.5.4 which is an example from VG, and the title has been labeled *health politics*. Example 4.5.5 is also an example from VG. In conjunction with Example 4.5.4 above, Example 4.5.5 is also *health politics*. Where FHI is short for *Folkehelseinstituttet*, which is Norway's national competence institution for public health, and KS stands for *Kommunesektorens Organisasjon*, which is the municipal sector's interest and employer organization.

Example 4.5.4 –

FHI: Nå ser vi kanskje at tiltakene gir større belastning enn sykdommen i seg selv

FHI: - Now we may see that the measures give greater burden than the disease itself

Example 4.5.5 –

KS: Krever ekstra satsing på distriktslegene

KS: Requires extra investment on doctors working in the districts

True news		Satirical	
Feature	Translation	Feature	Translation
ukraina	ukraine	siste	last/latest
trøndelag	trøndelag	5080	5080
strømstøtte	energy support	forskning	research
anmeldelse	review	lanserer	launches
innlandet	innlandet	venstre	political party
studenter	students	5080s	5080s
eurovision	eurovision	livets	of life
kritikk	criticism	lov	law
raymond	raymond	innfører	introduces
videre	further	sier	say

Table 4.9: Top 10 most informative features for Task 1 for true news and satire using a linear SVM.

4.5.2 SVM

Figure 4.3 presents the calculated confusion matrix for SVM for Task 1 on the test set. It shows that the model correctly identified 956 titles (468 as TP, 488 as TN), and misclassified 309 titles (128 as FN, and 181 as FP). In contrast to the confusion matrix for Naïve Bayes above, the SVM model seems to find it harder to classify regular news as regular news. The SVM model has wrongly classified 181 titles for being satire, when they in reality are true news. The model might misinterpret the titles, since satire mimics regular news, and the ambiguity of the headlines get miscalculated.

Table 4.9 presents the top 10 most informative features using SVM. The words are sorted by their score from top to bottom. For the satirical features, the source name *5080* appears two times. In addition, *siste* (last/latest) also occur for SVM as well as for Naïve Bayes, and can be correlated to the fact that some headlines start with *SISTE NYTT*, translated to *LATEST NEWS*, which is common for the satirical headlines, ref Figure 4.2. Again for the true news, some words cover international events, such as *Ukraina* (Ukraine) and *Eurovision*. Similarly to Naïve Bayes, some words also refers to national matters, like *Trøndelag*, *Innlandet*, *Raymond*, *strømstøtte* (energy support), and *studenter* (students).

4.5.3 Logistic Regression

Figure 4.4 presents the confusion matrix for logistic regression for Task 1 on the test set. It shows that the model correctly classified 954 titles (467 correctly classified as satire (TP), and 487 correctly classified as true news (TN)), and wrongly classified 311 titles (129 as true news, but was satire (FN), and 182 as satire, but was true news (FP). Comparable to the SVM's confusion matrix, the logistic regression model also finds

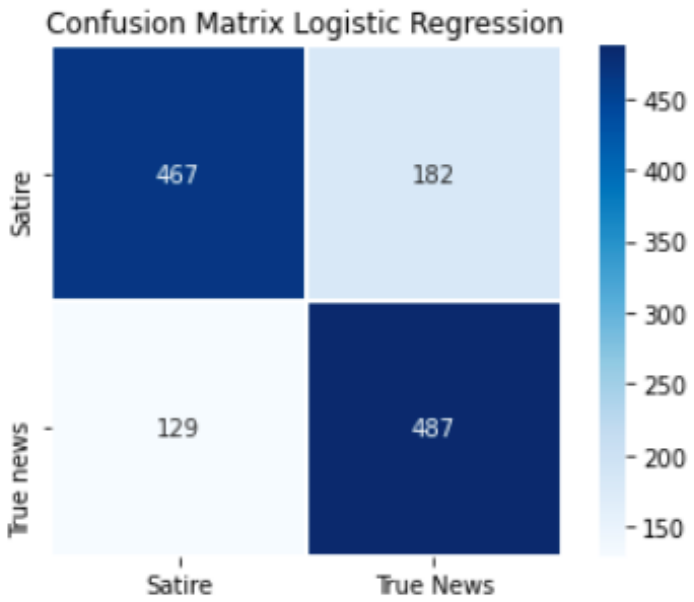

Figure 4.4: Confusion matrix for Task 1 using Logistic Regression.

it more difficult to distinguish regular news as regular news. The logistic regression model has wrongly classified 182 title for being satire (FP), when they are true news. Likewise as for SVM model above, the logistic regression model might misinterpret the titles, since satire mimics regular news, and the ambiguity of the headlines get miscalculated.

Table 4.10 represents the top 10 most informative features calculated by logistic regression. The true news is presented on the left, and satirical news on the right, and the words are sorted by their score from top to bottom. For the true news, 4 out of 10 features refer to international affairs, such as *Ukraina* (Ukraine), *Eurovision*, *ol* (The Olympics), and *omikron* (omicron). It also covers national matters, such as *Innlandet* and *trafikkulykke* (traffic accident). For the satirical news, again, words such as *siste* (last/latest), *nytt* (news), and *5080* are the top 3 features.

4.5.4 True news

For the real news, 84 titles were wrongly classified by all three models. Both NRK and VG had the same number of misclassification errors, with 42 titles each. When looking at the percentages of the misclassified titles in relation to the total number of articles in the test set, NRK has 13.29% of its titles misclassified, whereas VG has 14% of its titles misclassified.

The categories of *politics*, *health politics*, *agurknytt* (unimportant topics), and *general news* were the ones that were covered the most in the true news. As previously

True news		Satirical	
Feature	Translation	Feature	Translation
ukraina	ukraine	siste	last/latest
innlandet	innlandet	nytt	news
tre	three/tree	5080	5080
trafikkulykke	traffic accident	forskning	research
eurovision	eurovision	at	that
dette	this	sier	say
ol	the olympics	mann	man
omikron	omicron	norsk	norwegian
siktet	charged	år	year
videre	further	du	you

Table 4.10: Top 10 most informative features for Task 1 for true news and satire using Logistic Regression.

mentioned, to assign genres to the news each title was manually gone through and labeled by three people.

It's interesting to note that one of the titles in both the genuine and satirical categories was mislabeled as belonging to the other category when it concerned the Nobel Peace Prize.

When reading the headlines in the Table 4.11 below, it can be challenging to understand why the headlines were classified as satirical when they are, in fact, true news titles. Five of the titles came from NRK, while four of them were from VG. The most represented genres were *economics* and *general news*, with three titles dedicated to each subject area. Additionally, one title was devoted to each of the topics of *health politics*, *crime*, and *the Nobel Peace Prize*.

Two of the headlines labeled as economics, can have been misclassified as satirical because the models may see it as exaggeration. Such exaggerated numbers may seem unrealistic, and due to this, the models classify them as satirical. In addition, the last headline regarding 109 out of 194 students failing on their exam, may be viewed as exaggeration or absurdity, due to high numbers of students failing their class.

The other headlines can be more difficult to interpret why the headlines were wrongly classified. It can be due to single words, or how the sentences are built up. In the first headline it can be the word *frykter* (fears) being the triggering cause, but it is difficult to conclude on.

The titles shown in Table 4.12 demonstrate six headlines all incorrectly categorized by the ML models. The titles are examples of headlines where it can be difficult to determine if it is satire or true news. Showing that detecting figurative language can be challenging for humans, as well as for computers.

Four of the six titles presented in Table 4.12 have been labeled as *agurknytt* (unim-

True news - Norwegian	True news - Translated
Familien frykter Sverre Solli (96) skal dø mens han venter på saksbehandling i Karmøy kommune	The family fears Sverre Solli (96) will die while he waits for case processing in Karmøy municipality
Gjennomførte tilsyn hos fosterfar visste ikke om overgrepstiltale	Carried out supervision of foster father - did not know about abuse charges
Kutter nettleia med 4 milliarder i år	Cuts grid rent by 4 billion this year
Usikkert når det blir nytt billett-system	Uncertain when there will be a new ticket system
Reddet naboen ut av boligbrann: Vi var der til akkurat rett tid	Rescued the neighbor from a house fire: - We were there at just the right time
Statkraft firedobla resultatet vil dele ut 10 mrd	Statkraft quadruples the result - will distribute 10 billion
Årets fredspriskandidater snart klare	This year's Peace Prize candidates are soon ready
Har ikke gjort funn i Numedalslågen	Has not made any discoveries in Numedalslågen
109 av 194 strøk: Nå får hele kullet tilbud om omsensur	109 out of 194 failed - Now the whole class is offered circumcision

Table 4.11: Titles predicted as satire, true label is true news.

porant topics). The other two, have been labeled as *crime* and *health politics*. Respectively four of the headlines were from NRK, and two of them from VG. Two of the titles labeled as *agurknytt* were from NRK, and the other two were from VG. It may seem that headlines labeled as *agurknytt* use satirical queues, such as humor, absurdity, and exaggeration.

Given that pirates often reside on the sea, the headline addressing the suspected pirate and not releasing him on the sea seems to use absurdity in the sense that it appears silly to release someone suspected of something criminal at sea. And the fact that the suspected person is a pirate, makes it even more absurd.

The headline regarding Øyvind, and his three stolen bikes may be interpreted as exaggeration, considering the fact that his bicycle has been stolen three times already.

In addition, the title regarding the old person demanding better banking services, also comes off as an exaggeration. By using the term; *I am old, not an idiot*, it comes off as an overstatement.

The headline on the use of ketchup in the SFO (after-school program) appears humorous given that many people consume ketchup on a regular basis and it would seem silly to argue whether or not the SFO should include ketchup. However, not everyone

True news - Norwegian	True news - Translated
Forelsket på nett: Hvordan kan folk påstå at det ikke er ekte?	In love online: - How can people claim that it is not real?
Mistenkt pirat tas med til Danmark: "Ikke forsvarlig å slippe ham løs på sjøen"	Suspected pirate brought to Denmark: "Not justifiable to release him at sea"
Krever bedre banktjenester: Jeg er gammel, ikke idiot	Demands better banking services: - I'm old, not an idiot
Øyvind har blitt frastjålet sykler 3 ganger	Øyvind has had his bikes stolen 3 times
Tar opp kampen med ketchup på SFO	Takes up the fight with ketchup at SFO
Har fått kontakt med turgåar	Has been in contact with hiker

Table 4.12: Ambiguous titles, predicted as satire, true label is true news.

may be aware of the relevance of removing ketchup from SFO due to its high sugar content.

4.5.5 Satire news

The source with the greatest number of misclassified titles for satirical news is 5080, with 72 titles out of 112. Vredens-Gnag, with 31 incorrectly categorized titles, is the second most common source of misclassified titles. Eavisa misclassified nine headlines, whereas Satiriks had none misclassified titles.

Since 5080 is the most represented source, with 1501 total articles and 297 of them in the test set, it is reasonable to believe that the majority of the misclassified titles would come from this source. However, examining the percentage for all sources, 24.24% of the titles in the test set for 5080 are incorrectly categorised. Whereas, Vredens-gnag has 27.93% incorrectly categorized titles for the test set, Eavisa has 3.81% incorrectly classified titles for the test set. No titles on Satiriks are misclassified. Making Vredes-Gnag the most represented source of misclassified titles according to percentages. With only five headlines, Satiriks is the least represented source in the test set.

Politics, *sports*, and *humor* appear most frequently in the misclassified satirical headlines. This matches *Rubin et al.* (2016) description about satire news covering the same subject matters as true news. It can also be observed in both satire news and real news that the Nobel Peace Prize was misclassified under an incorrect title for both classes. In addition, genres such as *general news*, *health*, *heath politic*, *weather*, *celebrity news*, and *entertainment* are also found in the satire headlines.

In addition to the use of figurative language in the titles, absurdity, exaggeration,

contradictions, and humor is also frequently used in the satirical headlines. To give an illustration, Example 4.5.6 refers to a satirical title containing absurdity. Saying that the US will drop 500.000 electric scooters over Iran, which is an illogical act to do. Example 4.5.7 refers to a satirical title using humor, stating that plane crashes are caused by depressed birds.

Example 4.5.6 –
USA slipper 500 000 el-sparkesykler over Iran
The US drops 500 000 electric scooters over Iran

Example 4.5.7 –
Flykrasj skyldes deprimerte fugler
Plane crash is caused by depressed birds

In addition, the use of contradictions is also present in the satirical headlines, *e.g.* Example 4.5.8 referring to vaccine opponents asking the authorities to create a vaccine against corona.

Example 4.5.8 –
Vaksinemotstandere ber myndighetene skynde seg å produsere en vaksine mot corona
Vaccine opponents are asking the authorities to hurry up and produce a vaccine against corona

Humor is also applied in satire. The two headlines below being examples of the use of humor. Example 4.5.9 referring to a landslide accident, where a professor was taken by a flood of books, and Example 4.5.10 being about the US and EU freezing Putin's LinkedIn account.

Example 4.5.9 –
Skredulykke: Professor tatt av bokras
Landslide accident: Professor taken by a landslide of books

Example 4.5.10 –
USA og EU fryser Putins LinkedIn-konto
The US and EU freeze Putin's LinkedIn account

For the satirical news there was also some ambiguity. Where titles have been classified as true news, but they were satirical. It can be difficult to state why the ML models would classify some articles to be true news and others to be satire news. Looking

Satirical news - Norwegian	Satirical news - Translated
Megler fikk solgt bolig	Real estate agent sold home
Høstens nye realityTV-konsepter avslørt	This autumn's new reality TV concepts revealed
KOMMENTAR: Derfor raste taket på Sentrum Scene	COMMENT: The reason of why the roof of Sentrum Scene collapsed
Vi må tørre å tro på at vi vant i går	We must dare to believe that we won yesterday
Påbud om V-stil i alpint	Order for V-style in alpine
Vi kommer sterkt tilbake neste sesong	We will be back stronger next season
Vanvittig pengebruk i Bergen kommune	Insane spending in Bergen municipality
Tybring-Gjedde observert på Dombås	Tybring-Gjedde observed at Dombås

Table 4.13: Ambiguous titles, predicted as true news, true label is satire

at Table 4.13 with respect to the misclassified titles below, all headlines may be interpreted as real news, as they state affairs that are not unlikely to be written about in the news.

For instance, the first headline refers to a real-estate broker who sold a home, which is a common occurrence. In addition, the title about the reason why the roof at Sentrum Scene collapsed, is an incident that could have happened. Similarly for the other titles, when not having more context, it is possible to believe that the titles are authentic.

4.6 Summary of findings Task 1

Taking a look at Figures 4.1, 4.3, and 4.4 regarding the confusion matrices for Task 1 for the test set. All three ML models are similar in their calculations, with 314 (Naïve Bayes), 309 (SVM), and 311 (logistic regression) misclassified titles each.

In addition, looking at the misclassified headlines for both satire news and true news, it can be difficult to point out why the ML models wrongly classify some headlines, and correctly classify others. Looking at examples provided in Tables 4.11, 4.12, and 4.13, many of the titles can look similar for both types of input. With some headlines being easy to interpret as satire by humans even though the ML models classified them wrong. And other headlines being difficult to interpret due to the ambiguity of the title.

Looking at Tables 4.10 and 4.9, both the SVM and logistic regression models share more similar features than Naïve Bayes (see Table 4.8). This may be due to the fact that SVM and logistic regression are both binary classifiers, whereas Naïve Bayes is

a generative classifier. Looking at the top 4 features for SVM and logistic regression for the satirical news, 3 out of 4 features are the same, respectively *siste* (last/latest), *5080*, and *forskning* (research). Furthermore, they also have the feature *sier* (says) in common, which is ranked as the 10th most informative feature for SVM, and as 6th for logistic regression. The only feature all three models have in common for the satirical news is *siste* (last/latest), which is ranked 5th for Naïve Bayes, and 1st for both SVM and logistic regression.

For the most informative features for the true news, SVM and logistic regression share three similar features, *Ukraina* (Ukraine), *Innlandet*, and *videre* (further). Where *Ukraina* (Ukraine) and *videre* (further) share the same spot, respectively 1st and 10th (see Tables 4.9 and 4.10). Looking at Table 4.8 and Table 4.9, Naïve Bayes and SVM only have one feature in common for the true news, *anmeldelse* (review). Notably, Naïve Bayes and logistic regression do not share any features alike for the true news.

4.7 Task 2 – test set

The results from Task 2 – text as input on the test are reported in the following section. The confusion matrices of each of the three different models will be presented, in addition to the 10 most informative features. In the end there will be a brief summary of the most eminent findings for Task 2.

4.7.1 Naïve Bayes

Figure 4.5 presents the calculated confusion matrix for Naïve Bayes. It shows that the model correctly classified 1189 articles (636 as TP, and 553 as TN). For the misclassified, Naïve Bayes misclassified 76 articles. 63 of the articles were predicted as true news, when in reality they are satire (FN). 13 of the articles were wrongly classified as being satire, when they were true news (FP). From Figure 4.5 one can see that Naïve Bayes wrongly classifies more satire articles to be true news (FN), then true news articles to be satire (FP).

Table 4.14 presents the top 10 most informative features. The words are sorted by their score from top to bottom. The true news are represented to the left, and the satire news to the right. Naïve Bayes have predicted only numbers as being the most informative features for true news. For the satire news, the most informative features is stop words, which is similar to the top features for titles (see Table 4.8). This indicates that for the Naïve Bayes classifier, the most informative features are words that appear multiple times in an article, regardless of the meaning of the word.

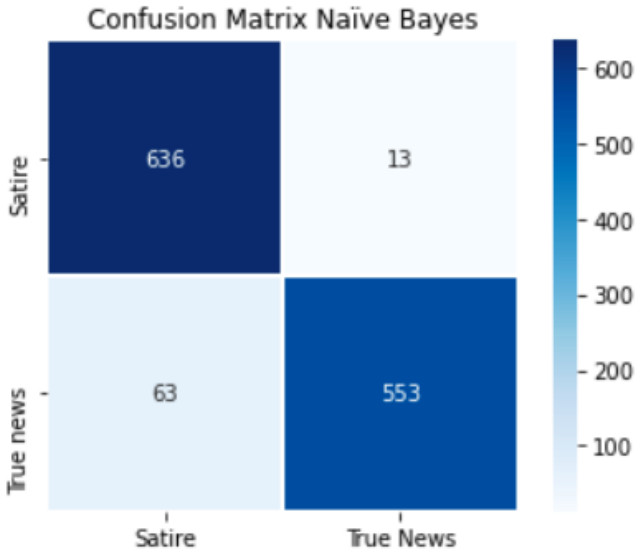

Figure 4.5: Confusion matrix for Task 2 using Naïve Bayes.

True News		Satirical	
Feature	Translation	Feature	Translation
105	105	er	is
1938	1938	og	and
1967	1967	det	the
1975	1975	at	that
1976	1976	på	on
2024	2024	som	as
213	213	en	one
270	270	har	has/have
3500	3500	jeg	I
360	360	til	to

Table 4.14: Top 10 most informative features for true news and satire, Task 2 using Naïve Bayes.

4.7.2 SVM

The SVM model has correctly classified 1241 articles (640 as TP, and 601 TN). SVM is the model with the lowest number of misclassified articles, with only 24 articles wrongly labeled (15 as FN, and 9 as FP). Same as for Naïve Bayes model, the SVM model has also misclassified more articles to be true news when they originally are satire (FN), then it has wrongly classified true news to be satire (FP).

The most informative features for SVM are presented in Table 4.15. For the true news, both the source names, *VG* and *NRK*, where the true news is gathered from, are presented in the top 10 features. This indicates that when the source name is present in the text, it may make it easier for the model in classifying it as true news. Respectively, the top 10 features for the satire news, both *5080* and *Eavisa* are present. In addition,

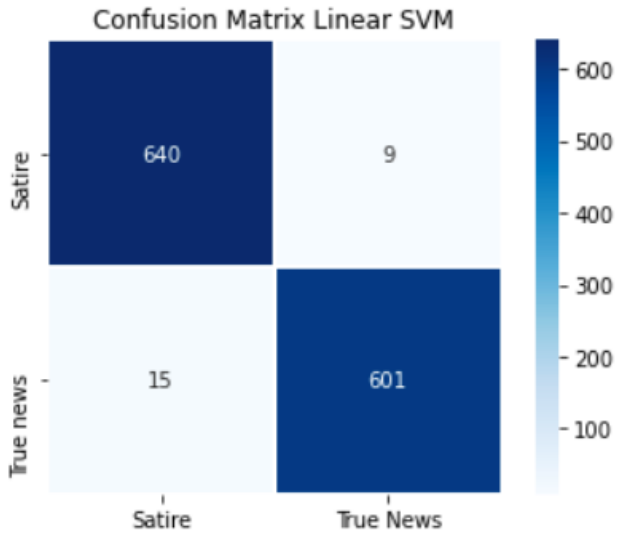

Figure 4.6: Confusion matrix for Task 2 using linear SVM.

the word *5080posten* (5080post) also appears in the top 10 features for satire. Which also refers to the source of 5080.

Similarly to the top 10 features for SVM on Task 1, the word *Ukraina* (Ukraine) also occurs for Task 2. Features such as *januar* (January), *mai* (May) may appear frequent due to fact that the first export of data was done in May 2021 and the second export in February 2022, and that the articles gathered was written in the month/months prior to the export, or in the same month. Words such as *oppdatert* (updated), *foto* (photo) and *publisert* (published) also occur frequently. This can be due to the fact that news articles contain the date when it was published, and the name of the person who has taken the photos. Additionally, the articles often get updated when new information needs to be brought up to date. Since the source names are some of the most informative features, indicating that the classifiers learn which articles belong to which publication source, and classify the articles accordingly.

4.7.3 Logistic Regression

For logistic regression, the model correctly classified 1214 articles (629 as TP, and 585 as TN). For the misclassified, 51 of the articles were wrongly labeled. 31 was wrongly classified as being true news, when by origin being satire (FN). For the FP, 20 articles were incorrectly classified as being satire, when in reality being true news. Similarly as for the Naïve Bayes model and the SVM model, logistic regression also produced more FN than FP. This can suggest that an article is more challenging to classify as satire, than it is to classify an article as true news. However, it is worth mentioning that logistic regression and SVM had a more equal distribution of misclassification than

True News		Satirical	
Feature	Translation	Feature	Translation
foto	photo	5080	5080
publisert	published	bare	just
januar	january	vi	we
skriver	write	eavisa	eavisa
oppdatert	updated	ja	yes
vg	vg	alle	all
nrk	nrk	5080posten	5080post
ukraina	ukraine	jo	yes
ntb	ntb	forteller	says
mai	may	nei	no

Table 4.15: Top 10 most informative features for true news and satire, Task 2 using a linear SVM.

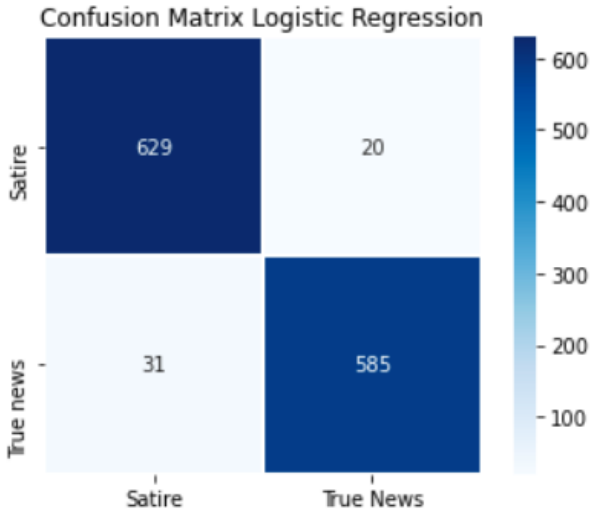

Figure 4.7: Confusion matrix for Task 2 using Logistic Regression.

Naïve Bayes. Indicating that Naïve Bayes struggles more in classifying the articles correctly than SVM and logistic regression.

Table 4.16 introduces the top 10 most informative features for logistic regression. Similarly to the top features for SVM on true news, words such as *publisert* (published), *foto* (photo), *januar* (January), *VG*, *oppdatert* (updated), *NRK*, *Ukraina* (Ukraine) also appears frequent for logistic regression. For the true news, SVM and logistic regression have 7 out of 10 words in common.

Looking at the most informative features for the satirical news, again words such as *5080* and *Eavisa* occur, implying that having the source name in the text helps classify the articles. This is same for true news where *VG* and *NRK* also are on the top 10 most informative features for logistic regression. Indicating that the models may learn ML domain specific features instead of satire specific features.

True News		Satirical	
Feature	Translation	Feature	Translation
publisert	published	5080	5080
foto	photo	vi	we
januar	january	bare	just
vg	vg	alle	all
oppdatert	updated	at	that
nrk	nrk	ja	yes
skriver	write	jo	yes
ukraina	ukraine	eavisa	eavisa
politiet	police	jeg	I
ntb	ntb	forteller	says

Table 4.16: Top 10 most informative features for true news and satire, Task 2 using Logistic Regression.

The satirical news share 8 out of 10 features, with the exceptions being *nei* (no) and *at* (that). This suggests that both SVM and logistic regression values the same set of words.

4.8 Summary of findings Task 2

All three models correctly classifies over 1000 articles, with SVM being the model that correctly classifies most articles, with 1241 correctly classified. Logistic regression correctly classifies 1214 articles, whereas Naïve Bayes have the least with 1189 correctly classified. Similar for all three models it seems that they find it harder to classify articles as satire, than it is to classify articles as true news, which is the opposite of Task 1.

Looking at the most informative features for all three models, one can see that Naïve Bayes distinguishes itself from SVM and logistic regression. Viewing Table 4.14 all the features for the true news are numbers, whereas none of the features for SVM and logistic regression have numbers for the true news. For the satirical features, Naïve Bayes have one similar feature word with logistic regression, being the word *at* (that). Whereas both SVM and logistic regression both share similar features for both true news and satire, with 7 out of 10 alike features for true news, and 8 out of 10 similar features for satire news.

4.9 Task 3 – test set

This section presents the error analysis for Task 3 on the test set. We present the confusion matrix for each model, as well as the top 10 most informative features. In the end,

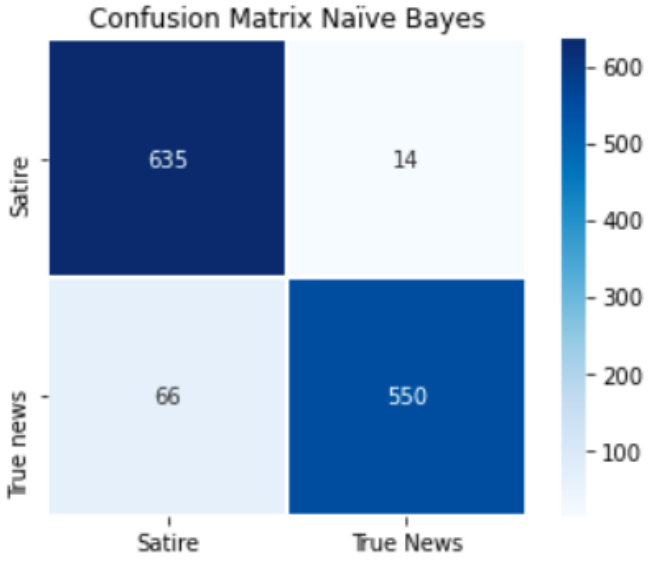

Figure 4.8: Confusion matrix for Task 3 using Naïve Bayes.

we give a short summary of the most important findings for Task 3.

4.9.1 Naïve Bayes

Figure 4.8 presents the confusion matrix calculated by Naïve Bayes for Task 3. It shows that the model has correctly classified 1185 full articles, including both title and text (635 as TP, and 550 as TN). For the wrongly classified articles, 66 were classified as FN (predicted true news, but is satire), and 14 as FP (predicted satire, but is true news). From the numbers deriving out of the confusion matrix, it is possible to see that the Naïve Bayes model find it harder in classifying satire than true news. From the misclassified, there is a significant difference in the FN's and the FP's. It is therefore fair to assume that Naïve Bayes finds it harder distinguishing if an article is satire, then distinguishing if an article is true news.

The top 10 most informative features for Task 3 using Naïve Bayes are presented in Table 4.17. The most informative features for Task 3 are almost identical to the most informative features for Task 2 (Table 4.14). The only difference can be seen in the satirical features, where *til* (to) and *jeg* (I) are listed in a different order. Nevertheless, all numbers and words are the same for both Tasks. This may be because the article's text has a greater impact on the model than the title has, since text as input reaches higher scores than using only title as input.

True News		Satirical	
Feature	Translation	Feature	Translation
105	105	er	is
1938	1938	og	and
1967	1967	det	the
1975	1975	at	that
1976	1976	på	on
2024	2024	som	as
213	213	en	one
270	270	har	has/have
3500	3500	til	to
360	360	jeg	I

Table 4.17: Top 10 most informative features for true news and satire, Task 3 using Naïve Bayes.

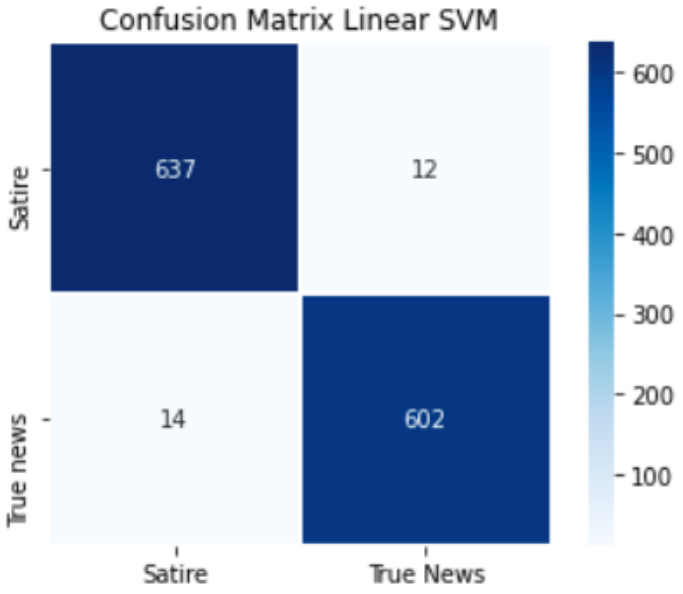

Figure 4.9: Confusion matrix for Task 3 using SVM.

4.9.2 SVM

Figure 4.9 presents the confusion matrix with the use of SVM for Task 3. SVM has correctly classified 1239 full articles, with 637 as TP, and 602 as TN. For the misclassified, SVM has only incorrectly labeled 26 full articles. The wrongly classified is almost equally distributed for the FN and FP. With 14 full articles being classified as FN, and 12 as FP. This is different from the prediction made by Naïve Bayes above, where there was an overweight of FN.

Table 4.18 gives an overview of the top 10 most informative features for Task 3 with SVM. Similarly, the most informative features for SVM are almost identical to the top 10 features for Task 2 (see Table 4.15). The difference can be seen in the order of the

True News		Satirical	
Feature	Translation	Feature	Translation
foto	photo	5080	5080
publisert	published	bare	just
januar	january	vi	we
oppdatert	updated	eavisa	eavisa
skriver	write	ja	yes
vg	vg	alle	all
nrk	nrk	jo	yes
ukraina	ukraine	5080posten	5080post
ntb	ntb	forteller	says
prosent	percent	nei	no

Table 4.18: Top 10 most informative features for true news and satire, Task 3 using SVM

words. For the satirical features, *jo* (yes), and *5080posten* (5080post) have switched places. For the true news, the features *oppdatert* (updated) and *skriver* (write) have a different order. In addition, the last words for the true news are different. Whereas for Task 2, the word was *mai* (Mai), and for Task 3, the word is *prosent* (percent).

4.9.3 Logistic Regression

Figure 4.10 presents the confusion matrix calculated for logistic regression on Task 3. From the figure, one can see that the model has correctly classified 1207 full articles, with 625 as TP, and 582 as TN. For the misclassified, logistic regression have wrongly predicted 58 full articles, with 34 FN and 24 FP. In accordance with the models above, logistic regression also wrongly classifies more articles to be true news, when they are satirical. This implies that all three models struggle more with predicting if an article is satire, than it does predicting if it is true news.

The top 10 most informative features for logistic regression is presented in Table 4.19. For the true news, all features are the same for Task 3 as for Task 2 using logistic regression (see Table 4.16). The only difference can be seen in the order of the features. The top feature is different for Task 3, with *foto* (photo) being number one. For Task 2 with logistic regression, the top feature was *publisert* (published), followed by *foto* (photo). In addition, the order of *oppdatert* (updated) and *VG* is different, as well as *Ukraina* (Ukraine) and *skriver* (write). For the satire news, the order of *at* (that) and *alle* (all) is different. Additionally, the satirical news has one different feature for Task 3, then Task 2. With the word *sier* (says) only appears to be a top 10 feature for Task 3.

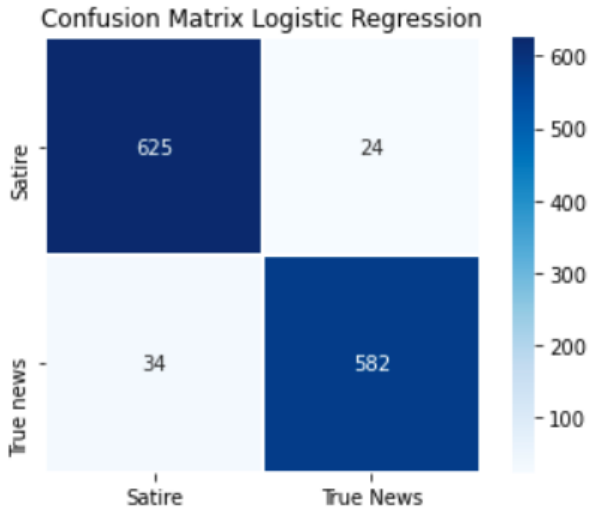

Figure 4.10: Confusion matrix for Task 3 using Logistic Regression.

True News		Satirical	
Feature	Translation	Feature	Translation
foto	photo	5080	5080
publisert	published	vi	we
januar	january	bare	just
oppdatert	updated	at	as
vg	vg	alle	all
nrk	nrk	ja	yes
ukraina	ukraine	jo	yes
skriver	write	eavisa	eavisa
politiet	police	sier	says
ntb	ntb	forteller	says

Table 4.19: Top 10 most informative features for true news and satire, Task 3 using Logistic Regression.

4.10 Summary of findings task 3

All three models correctly classified over 1000 full articles. SVM had the highest number of correctly classified (1239), in addition to the lowest number of wrongly classified (26). Second is logistic regression, with 1207 correctly classified, and 58 wrongly classified. Lastly, Naïve Bayes had 1185 correctly classified, and 80 wrongly classified. Equivalently for all three models, they struggled most in classifying the satire news correctly. This implies that the models find it harder to distinguish if an article is satire, then distinguishing if an article is true news.

As previously stated, the top 10 most informative features for Task 3, is similar to the top 10 features for Task 2 within each model. This indicates that the words that appear in the text of an article play a bigger role than the words that appear in the title.

It is difficult to conclude that this is the case, and it needs to be researched further.

4.11 Differences between title and text

The most informative features can help shed light to differences in titles and texts. The most informative features for the titles indicates more what the article text will be about. For instance, titles uses words such as *angrep* (attack), *anmeldelse* (review), *avlyse* (cancel), *strømstøtte* (energy support), *studenter* (students), *eurovision* (eurovision), *trafikkulykke* (traffic accident), *OL* (The Olympics), and *omikron* (omicron). The words presented above are excerpts of features from Naïve Bayes, SVM, and logistic regression for regular news (see Tables 4.8, 4.9, and 4.10). For the satirical news headlines, all features from the Naïve Bayes model just contained stopwords (see Table 4.8). For SVM and logistic regression, the features consisted of words such as *forskning* (research), *lanserer* (launches), *lov* (law), and *innfører* (introduces) (see Tables 4.8 and 4.9). This gives an illustration of words used in titles, for both regular news and satire news that allegedly expresses the content of the texts.

By contrast, the most informative features for text differ from title in the sense that one does not get too much information out of the text features about what the content of the text might be, as one does for title features. The text features for Naïve Bayes, only consists of numbers for the true news (for both Task 2 and Task 3). Furthermore, similar to the most informative features for Naïve Bayes on titles, the features for satire news text are also mostly stopwords. In comparison, for SVM and logistic regression, words such as *foto* (photo), *publisert* (published), *oppdatert* (updated), *VG*, *NRK*, and *skriver* (write) appears, and gives little to none indication about what the article is about (see Tables 4.14, 4.15, 4.18, and 4.17). Demonstrating that the most informative features for true news for text says little about the text content. When looking at the most informative features for the satirical news it is mostly stopwords, in addition to source names, (*5080* and *Eavisa*) (see Tables 4.14, 4.15, 4.18, and 4.17).

To conclude, the most informative features for both title and text indicate that features from the titles give more insight into what the article will be about then text gives.

4.12 Comparison with related work

Table 4.20 shows our results compared to previous research done on satire detection. In general, more research has been carried out on the English language than for other languages. The results on English corpora have achieved F1-scores results between 79% (*Burfoot and Baldwin*, 2009) and 93% (*Frain and Wubben*, 2016). English satire de-

tection has also been done in a multi-language aspect (*Barbieri et al.*, 2015b), where they obtained a 76.3% F1-score for the English corpora. Moreover, methods on Turkish (*Onan and Toçoğlu*, 2020; *Toçoğlu and Onan*, 2019) also obtained good accuracy results with respectively 89% and 97.72%. In addition, the satire detection study on French (*Ionescu and Chifu*, 2021) also produced good accuracy results regarding text as input (97.48%). Furthermore, the study by *Casalino et al.* (2021) on Italian news, produced the best result with an F1-score of 98.9%. Similarly, satire detection for the Arabic language (*Saadany et al.*, 2020) produced only 0.41% lesser result with a F1-score of 98.49%. In general, the results are similar for every language, with the poorest result being for the German corpus (F1-score 66.5%) (*McHardy et al.*, 2019).

However, it is difficult to establish whether a work is better or worse than our proposal, since the methods, corpora, and languages are different. Therefore, we believe that comparing the various methodologies in the literature is challenging due to the various methods applied. Accordingly, the datasets used for each experiment differ significantly with regards to content, size, topics and language. In addition, satire can be culturally loaded, and the manifestation of satire in the language used might be more complex in some languages compared to others.

4.13 Limitations

Despite our models achieving good performance values, we are aware that out work has some limitations. When creating the TF-IDF vectorizer, the threshold (minimum document frequency) is set to be 5. This can have an opposite effect when it comes to running the models when title is used as input. When setting the threshold to be 5, the word must appear in at least 5 documents for it to be considered. Considering the title input does not contain many words to begin with, it can make words that are important to not be considered by the model because it only appears *i.e.,* one time.

The TF-IDF term counting does not take into consideration the context nor the word order. This oversimplification eliminates numerous nuances of human communication. Some phenomena, such as homonym, synonymy, and polysemy, making it nearly impossible to determine their meaning in the absence of context.

Another limitation of our work that we are aware of is that our dataset could be cleaned more. Stop words, punctuation, and numbers could have been removed to remove noise from the articles. Looking at the most informative features for Task 2 and Task 3 (Tables 4.14, 4.15, 4.16, 4.17, 4.18, and 4.19), source names such as *VG*, *NRK*, *5080*, and *Eavisa* may contribute to help the models distinguish between if an article is satire or regular news. Suggesting that this enables the classifier to learn

which article belongs to which genre, and classify the articles accordingly. In addition, we could have set a `max_df` in the vectorizer in order to remove corpus-specific stop words based on intra corpus document frequencies.

Moreover, having the same source in the train, validation, and test sets may also cause the classifiers to learn domain-specific features rather than satire-specific features. The models are trained using the training set's features. Which causes the ML models to learn characteristics included in the training set's sources. Due to learning domain-specific features in the training set, when the ML models are executed on the validation set and the test set containing the same sources as the training set, the ML models may know which features to look for.

Due to limited time, we were not able to further investigate the misclassifications of Tasks 2 and 3, *i.e.* using text at input, and at the combination of title and text as input. It clearly would have taken much more time to manually annotate and analyse each misclassification, and identifying words and sentences that would distinguish the articles from being satire and regular news in the time accessible. However, the top 10 most informative features for Task 2 and Task 3 give some insights into why articles have been correctly or incorrectly classified. As mentioned above, words such as *VG*, *NRK*, *5080*, and *Eavisa* may not have been present in the articles misclassified. This is only an assumption, and needs to be researched further.

In this thesis we only look at the top 10 most informative features. This is also a limitation. It would have been possible to increase this number and look at the top 100 or top 200 features, and semi-automatically analysed them. This would have given us better insight into the overlap between the features of each model, and give us more knowledge about the most important features to each of the ML models.

	Language	Model	F-measure/accuracy
Burfoot and Baldwin (2009)	English	SVM + BNS	79.8%
Barbieri et al. (2015a)	Spanish	SVM + 7 features	81.4%
Barbieri et al (2015b)	English Spanish Italian	SVM + language independent features	76.3% 81.6% 80%
Frain and Wubben (2016)	English	SVM + BoW, 8 textual features unigrams	93%
Rubin et al. (2016)	English	SVM + 5 features	87%
Goldwasser and Zhang (2016)	English	COMSENSE	80.8%
Reganti and Bajpai (2016)	English	Ensemble classifier + 7 features	Product reviews 77.96% Twitter 78.16% Newswire 79.02%
Yang et al. (2017)	English	4-Level Hierarchical Network	91.46%
del Pilar Salas-Zárate et al. (2017)	Mexican Spanish	SMO	85.5% 84%
De Sarkar et al. (2018)	English	CNN + syntactic information	91.59%
McHardy et al. (2019)	German	word2vec + neural networkd	66.5%
Toçoglu and Onan (2019)	Turkish	SVM + unigrams	89%
Onan and Toçoglu (2020)	Turkish	neural networks + GloVe	97.72%
Apolinario-Arzube et al. (2020)	European Spanish Mexican Spanish Full dataset	FastText + BiGru	83.523% 91.431% 85.838%
Saadany et al. (2020)	Arabic	CNN + FastText	98.49%
Li et al. (2020)	English	ViLBERT	Text+Images 92.16%
Rogoz et al. (2021)	Romanian	Char-CNN	71.09%
Ionescu and Chifu (2021)	French	CamemBERT + DA	Text 97.48% Title 74.07%
Casalino et al. (2021)	Italian	sAttBLSTMConvNet	98.9%
Our approach	Norwegian	SVM + TF-IDF	Text 98% Title 76% Title and text 98%

Table 4.20: Comparison with related work.

Chapter 5

Conclusion and Future Work

In this thesis we present the first attempt to automatically detect satire in Norwegian news articles. We trained and developed three classification methods, namely Naïve Bayes, SVM, and logistic regression, based on TF-IDF feature weights. All three ML models achieved similar results using the same type of inputs. We use a dataset of both legitimate and satirical Norwegian news sites that contain news articles in genres such as politics, economics, health, and health politics. In total, it incorporates 6322 articles, whereas half (3161) are satirical. Using this corpus we performed three satire detection tasks using different input representations: (i) considering only the use of headlines, (ii) considering only the use of article texts, and lastly (iii) considering both titles and texts.

Our main findings result in a set of observations about what types of input that was most and least useful in detecting satire in a Norwegian corpus. After training and evaluating the ML models on the test set, we achieved the top accuracy score of 98% using text as input using SVM, as well as a 98% accuracy score on the combination of both titles and texts as input using SVM.

We also observe that satire detection on news headlines was significantly more challenging. The top accuracy score being 76% with SVM (see table 5.1). The results for the titles provide similar result as *Burfoot and Baldwin* (2009) and *Ionescu and Chifu* (2021), with *Burfoot and Baldwin* (2009) reaching an overall F1-score of 79.8% and *Ionescu and Chifu* (2021) reaching an accuracy score of 74.07% on titles. We therefore conclude that the task of automatically detecting satire using only headlines is quite challenging. We believe that this is due to lack of context and the short nature of titles, where classification models have too little information to pick up enough signal for classification. Some of the strategies to this limitations we have discussed in Chapter 4 is to remove any lower bound for word frequencies when creating input vectors for the task using only title as input data.

Input	Accuracy score	ML model
Title	76%	SVM
Text	98%	SVM
Title + Text	98%	SVM

Table 5.1: The best results obtained for each input type.

The purpose of this research was to investigate satire detection for a Norwegian corpus. Our work therefore aimed at answering the following research questions:

- **RQ1**: How will simple machine learning models perform in satire detection for Norwegian?
- **RQ2**: Which types of input achieve the highest classification scores?
- **RQ3**: Which aspects of satire are difficult to handle by simple machine learning models?

Research Question 1 We measured the classification performance of three simple ML models (Naïve Bayes, SVM, and logistic regression). From Tables 4.4, 4.5, and 4.6, simple ML models seem to work well on a Norwegian corpus, and our results are similar to previous studies for other languages, especially English (Table 4.20).

Research Question 2 Our models indicate that having more text achieves higher classification. When comparing the results accomplished by our models with each type of input, one can see that using text, and a combination of title and text obtains higher classification than when using only title as input. Observing the results from Task2 and Task 3, the best accuracy score is the same (98%) for both. However, the result on Task 2 is slightly better considering the Precision score for true news (99%), and the Recall for satire (99%). Our models seem to reach higher accuracy values when using texts and titles as input, compared to other works combining these two types of input. The only exceptions are *Casalino et al.* (2021) (98.9%) and *Saadany et al.* (2020) (98.49%).

Research Question 3 Since the models achieve high F1-score and accuracy scores for text as input, in addition to full articles containing both title and text, it seems that the aspects that are difficult for our simple machine learning models are shorter texts. Also, looking at the confusion matrices for titles (Figures 4.1, 4.3, and 4.4), we observe that the models find it harder to determine if a title is regular news rather than satire, *i.e.*, regular news are more often misclassified. This observation does not hold for the two other Tasks, 2 and 3, as the opposite holds for text as input and title+text as input.

There, the ML models find it harder to distinguish satire from true news (see Figures 4.5 - 4.10), *i.e.* satire is more often misclassified.

5.1 Contributions

Our main contributions are to the advancement of research on satire classification for Norwegian online articles. Creating and evaluating three machine learning methods (Naïve Bayes, SVM, and logistic regression), and analysing which aspects of satire classification are difficult to handle by simple machine learning models.

Collected a Norwegian corpus We introduce a dataset consisting of Norwegian news articles collected from two regular and four satirical news sources (VG, NRK, 5080, Eavisa, Vredens-Gnag, and Satiriks), which allowed us to perform cross-domain satire detection on three different types of input: (i) Task 1 – title as input, (ii) Task 2 – article text as input, and (iii) Task 3 – a combination of title and text. In addition, the dataset introduced consists of an equal distribution of satirical and non-satirical articles from different sources with various topic domains.

Working models for satire classification for the Norwegian language We propose simple and effective supervised machine learning models for classifying articles as satirical or regular news. Our models can take different types of inputs, from short text sequences (as titles), to longer sequences (entire documents). To the best of our knowledge, this is the first attempt at automatically classifying news articles as satirical or regular for the Norwegian language.

Analysis of the ML models with various input We found that short text inputs made it more difficult for the ML models to detect satire compared to longer text input. The classification performance also went down when only the title was used as input. This we believe shows the importance of context, which can only be achieved with longer text sequences.

5.2 Future work

This work has targeted satire classification for Norwegian news articles, and gives several possibilities for further research. In order to enhance the performance of our models, one could further improve the preprocessing of our dataset by *e.g.* removing domain names, stop words, and numbers. This we believe can enable the models to learn satire-specific features rather than domain-specific features.

Another area to investigate, is to apply deep learning architectures for trying to solve the satire detection classification task. From research presented in chapter 2 one can see that both machine learning approaches and deep learning architectures achieve similar results. It would be interesting to look into this further, and to see if this applies to the Norwegian language as well.

In this study, we only took article texts and headlines into consideration. Another interesting research avenue would be to look into multimedia models that incorporate information both from texts and images. *Li et al.* (2020) produced good results for using text and corresponding visualization, which we believe can be mirrored for Norwegian as well.

Our error analyses has shed light on several interesting aspects of our satire detection models, and we believe that a further investigation into the missclassifications of models using entire texts or combinations of texts and titles will make the challenges of satire detection even clearer. Due to time constraints we were not able to carry out these analyses, but we think that they would be valuable next steps to take for satire detection for Norwegian.

By investigating all of the aforementioned research possibilities, we believe that many of our findings can be corroborated, and even more contributions and findings can extend the research on satire detection for Norwegian.

Bibliography

Apolinario-Arzube, Ó., J. A. García-Díaz, J. Medina-Moreira, H. Luna-Aveiga, and R. Valencia-García (2020), Comparing deep-learning architectures and traditional machine-learning approaches for satire identification in spanish tweets, *Mathematics*, *8*(11), 2075. 2.2.2, 2.3, 3

Barbieri, F., F. Ronzano, and H. Saggion (2015a), Is this tweet satirical? a computational approach for satire detection in spanish, *Procesamiento del Lenguaje Natural*, (55), 135–142. 2.1, 2.2.1

Barbieri, F., F. Ronzano, and H. Saggion (2015b), Do we criticise (and laugh) in the same way? automatic detection of multi-lingual satirical news in twitter, in *Twenty-Fourth International Joint Conference on Artificial Intelligence*. 2.2.1, 4.12

Burfoot, C., and T. Baldwin (2009), Automatic satire detection: Are you having a laugh?, in *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pp. 161–164. 1.2, 2.2.1, 2.2.2, 3.6, 4.1.4, 4.12, 5

Casalino, G., A. Cuzzocrea, G. L. Bosco, M. Maiorana, G. Pilato, and D. Schicchi (2021), A novel approach for supporting italian satire detection through deep learning, in *International Conference on Flexible Query Answering Systems*, pp. 170–181, Springer. 2.2.2, 2.3, 4.12, 5

Colletta, L. (2009), Political satire and postmodern irony in the age of stephen colbert and jon stewart, *The Journal of Popular Culture*, *42*(5), 856–874. 2.1

De Sarkar, S., F. Yang, and A. Mukherjee (2018), Attending sentences to detect satirical fake news, in *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3371–3380, Association for Computational Linguistics, Santa Fe, New Mexico, USA. 1, 2.2.2, 2.3

del Pilar Salas-Zárate, M., M. A. Paredes-Valverde, M. Á. Rodriguez-García, R. Valencia-García, and G. Alor-Hernández (2017), Automatic detection of satire in

twitter: A psycholinguistic-based approach, *Knowledge-Based Systems*, *128*, 20–33. 2.1.2, 2.2.1, 4.5.1

Farrell, J. (2015), Politics: Echo chambers and false certainty, *Nature Climate Change*, *5*, 719–720, doi:10.1038/nclimate2732. 1

Frain, A., and S. Wubben (2016), SatiricLR: a language resource of satirical news articles, in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 4137–4140, European Language Resources Association (ELRA), Portorož, Slovenia. 2.1.1, 2.2.1, 2.3, 4.12

Gandomi, A., and M. Haider (2015), Beyond the hype: Big data concepts, methods, and analytics, *International Journal of Information Management*, *35*(2), 137–144, doi:https://doi.org/10.1016/j.ijinfomgt.2014.10.007. 2

García, S., J. Luengo, and F. Herrera (2015), *Data preprocessing in data mining*, vol. 72, vii pp., Springer. 3.3

Gilmore, J. T. (2017), Satire. 2.1

Goldwasser, D., and X. Zhang (2016), Understanding satirical articles using commonsense, *Transactions of the Association for Computational Linguistics*, *4*, 537–549. 2.2.2

Horne, B., and S. Adali (2017), This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news, *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), 759–766. 3.1

Howell, L. (2013), Digital wildfires in a hyperconnected world. (document), 1

Ionescu, R. T., and A. Chifu (2021), Fresada: A french satire data set for cross-domain satire detection, *CoRR*, *abs/2104.04828*. 2.2.2, 3.6, 4.1.4, 4.2.4, 4.12, 5

Jiang, M., Q. Gao, and J. Zhuang (2021), Reciprocal spreading and debunking processes of online misinformation: A new rumor spreading–debunking model with a case study, *Physica A: Statistical Mechanics and its Applications*, *565*, 125,572. 1

Jurafsky, D., and J. H. Martin (2020), *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, third edition draft ed. (document), 3, 3.5, 3.5, 3.5, 3.5, 3.5, 4.5.1

Kummervold, P. E., J. De la Rosa, F. Wetjen, and S. A. Brygfjeld (2021), Operationalizing a national digital library: The case for a norwegian transformer model, *arXiv preprint arXiv:2104.09617*. 1

Le, Q., and T. Mikolov (2014), Distributed representations of sentences and documents, *31st International Conference on Machine Learning, ICML 2014*, *4*. 2.2.2

Li, L., O. Levi, P. Hosseini, and D. A. Broniatowski (2020), A multi-modal method for satire detection using textual and visual cues, *arXiv preprint arXiv:2010.06671*. 2.2.2, 5.2

McHardy, R., H. Adel, and R. Klinger (2019), Adversarial training for satire detection: Controlling for confounding variables, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 660–665, Association for Computational Linguistics, Minneapolis, Minnesota, doi: 10.18653/v1/N19-1069. 2.2.2, 4.12

Müller, A. C., and S. Guido (2017), *Introduction to Machine Learning with Python*, 27, 58, 70-72, 269, 338-339 pp., O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol. 2.2.1, 3, 3.2.2, 3.4

Nyhan, B., and J. Reifler (2010), When corrections fail: The persistence of political misperceptions, *Political Behavior*, *32*(2), 303–330. 1

Onan, A., and M. A. Toçoğlu (2020), Satire identification in turkish news articles based on ensemble of classifiers, *Turkish Journal of Electrical Engineering & Computer Sciences*, *28*(2), 1086–1106. 2.1.1, 2.2.1, 2.2.2, 2.3, 4.12

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011), Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research*, *12*, 2825–2830. 3.2.2, 3.3, 3.4

Qi, P., Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning (2020), Stanza: A Python natural language processing toolkit for many human languages, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 3.2.1

Reganti, A. N., T. Maheshwari, U. Kumar, A. Das, and R. Bajpai (2016), Modeling satire in english text for automatic detection, in *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pp. 970–977, IEEE. 2.1.1, 2.2.1

Reyes, A., P. Rosso, and D. Buscaldi (2012), From humor recognition to irony detection: The figurative language of social media, *Data & Knowledge Engineering*, *74*, 1–12. 2.1.1

Rogoz, A.-C., M. Gaman, and R. T. Ionescu (2021), Saroco: Detecting satire in a novel romanian corpus of news articles, *arXiv preprint arXiv:2105.06456*. 2.2.2

Rubin, V., N. Conroy, Y. Chen, and S. Cornwell (2016), Fake news or truth? using satirical cues to detect potentially misleading news, in *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pp. 7–17, Association for Computational Linguistics, San Diego, California, doi:10.18653/v1/W16-0802. 1.2, 2.1, 2.1.2, 2.2.1, 2.2.2, 3.6, 4.1.4, 4.5.5

Rundell, M., L. Potter, K. Maxwell, and D. Nicholls (2022), Macmilliandictionary.com. 4.4

Saadany, H., C. Orasan, and E. Mohamed (2020), Fake or real? a study of Arabic satirical fake news, in *Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM)*, pp. 70–80, Association for Computational Linguistics, Barcelona, Spain (Online). 2.2.2, 2.3, 4.12, 5

Shin, J., L. Jian, K. Driscoll, and F. Bar (2018), The diffusion of misinformation on social media: Temporal pattern, message, and source, *Computers in Human Behavior*, *83*, 278–287. 1

Thrall, F. W., and A. Hibbard (1960), A handbook to literature, pp. 436–437, The Odyssey Press, Inc. 2.1

Toçoğlu, M. A., and A. Onan (2019), Satire detection in turkish news articles: A machine learning approach, in *Big Data Innovations and Applications*, edited by M. Younas, I. Awan, and S. Benbernou, pp. 107–117, Springer International Publishing, Cham. 2.2.1, 4.12

Vraga, E. K., and L. Bode (2020), Defining misinformation and understanding its bounded nature: Using expertise and evidence for describing misinformation, pp. 136–144, doi:10.1080/10584609.2020.1716500. 1

Yang, F., A. Mukherjee, and E. Dragut (2017), Satirical news detection and analysis using attention mechanism and linguistic features, in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1979–1989, Association for Computational Linguistics, Copenhagen, Denmark, doi:10.18653/v1/D17-1211. 2.2.2, 2.3, 3.2.1, 3.6

Yitzhaki, M. (2002), Relation of the title length of a journal article to the length of the article, *Scientometrics*, *54*(3), 435–447. 3.2.1

Zhang, W., G. Yang, Y. Lin, C. Ji, and M. M. Gupta (2018), On definition of deep learning, in *2018 World automation congress (WAC)*, pp. 1–5, IEEE. 2.2.2