Sentiment Analysis of Norwegian Twitter News Entities

Jon Atle Gulla*[1], John Arne Øye**[2], Xiaomeng Su***[3], and Özlem Özgöbek†[1]

[1]Department of Computer Science, NTNU, Trondheim, Norway
[2]Acando, Trondheim, Norway
[3]Department of Informatics and e-Learning, NTNU, Trondheim, Norway

Abstract. Microblogging websites like Twitter complement traditional news agencies and have become important sources of information about news events. In particular, aggregated sentiment values from Twitter news messages may tell us about the overall popularity of news entities or people's general perception of news events or entities. On the basis of a Norwegian Twitter news dataset we examine how the sentiments of Twitter news messages can be extracted using Naive Bayes, Support Vector Machines and Maximum Entropy. Our analysis also includes the use of linguistic features and lexical sentiments from SentiWordNet in an attempt to improve the accuracy of the techniques. The results show that there is some gain in including part-of-speech features or predefined sentiments from SentiWordNet. Support Vector Machines has the highest accuracy for both subjectivity classification and polarity classification, though the differences are small and all techniques' performances increase steadily with the size of the dataset. Moreover, our work demonstrates that sudden changes of news entity sentiments tend to be attributed to concrete entity-relevant news events.

1 Introduction

Over the last few years Twitter has become an important source of information about unfolding news events that have not yet been properly picked up by news agencies, or where news reporting is difficult or unreliable for some reason. Any owner of a Twitter account may post small messages - tweets - of up to 140 characters that can be widely distributed and read over the internet. In spite of the brevity of these messages, Twitter has been extensively used to report or comment on incidents all over the world, and Twitter now has more than 300 million active users and an excess of 500 million tweets per day.

These tweets may be posted by users that are present at the event and that have no particular experience with traditional news reporting. As a news source

* jag@idi.ntnu.no
** john.arne.oye@acando.no
*** xiaomeng.su@ntnu.no
† ozlemo@idi.ntnu.no

Twitter complement the publications of traditional news agencies in several interesting ways: 1) the news may be quickly and directly reported as the event takes place, 2) the news are often reported by people that are physically present or in other ways have a direct link to the event, 3) the aggregation of tweets from many users cancels out individual misconceptions and presents a possibly more reliable collective perception of the event, and 4) Twitter users provide other perspectives than professional journalists and may serve as a corrective to the news agencies that are driven by increasing time pressure and declining revenues.

There are however limitations with Twitter that renders it somewhat unsuitable as a general news source:

- The 140-character limitation makes it unfeasible to explain matters in sufficient detail.
- The lack of any other structural elements than hashtags and user references introduce ambiguities that are difficult to handle manually and very challenging computationally.
- Individual tweets may be wrong, contradictory, incomplete or misleading.
- There are too many tweets to read and no satisfactory way of selecting the most appropriate ones.

Even though individual tweets should not necessarily be trusted, the message formed by collective streams of news tweets carry more weight and may expose other types of information than what is conveyed in reports from conventional news channels. Since people use Twitter for posting opinions on a variety of news topics and express their attitudes towards products or people on a daily basis, it seems interesting to aggregate messages and try to extract a sense of general sentiment over time for particular news entities.

In the SmartMedia project at NTNU in Norway we are developing a news aggregator in collaboration with one of the largest media houses in Norway [9] [28] [12]. All the major newspapers in Norway are indexed as part of this mobile application, though we also include user-generated material from Twitter where this is appropriate. In particular, we make use of Twitter to analyze people's aggregated sentiment perception of important news entities over time.

We have in our project used and compared Naive Bayes, Maximum Entropy and Support Vector Machine for sentiment analysis of Norwegian news tweets. An annotated Norwegian data set has been employed, and a variety of feature sets have been compared for each technique. An important part of the work has been the extraction of lexical words - or concepts - from the news text that can be associated with an a priori sentiment from a sentiment ontology like SentiWordNet [5]. The analysis shows to what extent semantic enrichment from SentiWordNet can improve the quality of the classifications. In this paper, we also show the sentiments over some time of a particular news entity - the Norwegian prime minister - to demonstrate the relationship between sentiment patterns and related news events.

The rest of paper is structured as follows. In Section 2 we discuss the use of Twitter for posting information about news events. We assess the overall problem

of extracting sentiments from Twitter and also present the Twitter dataset that is used in our work. Section 3 introduces related work on news sentiment analysis, with a particular emphasis on the use of Naive Bayes, Support Vector Machines and Maximum Entropy. The whole sentiment analysis process is explained in Section 4. This includes the enrichment of the feature set with part-of-speech tags and predefined concept sentiments from SentiWordNet, and the configuration of the machine learning techniques for subjectivity classification and polarity classification. Whereas the overall results are presented and discussed in Section 5, the application of entity sentiments over time is briefly discussed in Section 6. The conclusions are given in Section 7.

2 Twitter News Data

Twitter, like many other social media platforms, allows people to express and distribute their views across geographical, national and social borders. The service was founded in 2006 and is today one of the largest microblogging services available. Apart from making it easier for people to communicate, these social media networks collect large amounts of data that can be aggregated and analyzed to identify for example breaking news as they are emerging. As a result social media networks have provided valuable information in real-time about crisis situations such as earthquakes and tsunamis [19]. [25] examined how earthquakes could be detected using Twitter. Their research, which regarded Twitter users as sensors and tweets as sensor data, suggest that up to 96% of the earthquakes with intensity 3 or more occurring in the examined area could be identified from the analysis of Twitter users. In [11], the authors show how the news of Osama Bin Laden' s death spread on Twitter before the mass media could get the news confirmed.

Fig. 1. Tweet from Norwegian prime minister.

Often, though, Twitter is used to express very personal attitudes or opinions about products, companies or people. The tweet in Figure 1 shows an example of this. The user is Erna Solberg, the prime minister of Norway, and she is thanking the previous prime minister for winning a prize for the best tweet in Norway in 2015: 'I would also like to thank you, @jensstoltenberg. If you had not lost your herring recipe, I would probably not have won this prize. #smd2014.' @jensstoltenberg is a reference to a particular user Jens Stoltenberg, who is

the previous prime minister of Norway. The hashtag #smd2014 refers to the Social Media Days conference in Norway in 2014. A tweet may also contain emoticons, like small smiley faces, but can normally not exceed 140 characters in total. In comparison, the average non-finance news article in Norwegian online newspapers is of about 220 words [9].

2.1 Sentiments of Tweets

Sentiments may be associated with individual words, phrases, sentences, paragraphs or documents. They express opinions of some entities in terms of positive, negative or neutral attitudes towards the entities. The entities may be like people or products, though they may also be components or aspects of higher-level entities. For example, the objective of a sentiment analysis task may be to assess the sentiments towards a company like Sony (brand reputation), but it may also be to extract people's opinions of Sony's mobile phones or even people's perception of the battery life of these phones. In general, [17] defines a sentiment (opinion) for an undecomposed target as follows:

A quadruple, (g,s,h,t), where g is the sentiment target, s is the sentiment about the target, h is the sentiment holder, and t is the time when the opinion was expressed.

Take a look at Figure 1. Analyzing the sentiments of this tweet, we see that Erna Solberg is the sentiment holder and the opinion was stated on February 5, 2014. It is more complicated to identify the target and estimate the sentiment about the target. The text is generally positive, as she is thanking a political opponent and seems to be happy to win an award. It would be tempting to conclude that the award #smd2014 is the target, but we need to keep in mind that this tweet is posted in a political context, and she is having fun with teasing the previous prime minister (Jens Stoltenberg) that had to move out of the prime minister residence and lost his recipe in the process. She is happy about the election, which is not directly discussed in the tweet, and this award is just mentioned to make the posting a bit childish and funny.

Analyzing the sentiments of news entities from Twitter is notoriously problematic for several reasons. In the first place, it is often difficult to identify the entity that forms the topic of a particular tweet. As seen above, the entity may not be directly mentioned in the text, or it is referred to indirectly by means of other entities that are somehow related. A second issue is the shortage of sentiment-carrying adjectives and adverbs in Tweets. Adjectives are normally very useful, as their sentiments do not change much from one domain to another and can be retrieved from sentiment lexica to calculate aggregated sentiments. Without sufficient adjectives and adverbs, you are left with analyzing context-dependent sentiment values of phrases, which diminishes the value of sentiment lexica in the analysis. Due to the nature of Twitter users, there may also be deliberate ambiguity or irony in tweets that affect the aggregation of sentiment values.

2.2 Norwegian News Data set from Twitter

In this work we have built a data set of Norwegian news tweets over a period of 30 days. To ensure that we had enough tweets for particular news entities, we collected the tweets using three separate strategies:

- Collect news tweets completely at random (33% of data set)
- Collect news tweets about the football club Rosenborg (31% of data set)
- Collect news tweets about prime minister Erna Solberg (36% of data set)

A summary of the total data set is given in Table 1. It contains 1847 Norwegian tweets, with an average of 16.1 words per tweet. There are a few emoticons like smiley faces in the dataset, but surprisingly few. The tweets were posted by 1,312 users, who themselves refer to 1,844 users in the texts. Simple word correction was performed on the tweets, as many users were deliberately using improper spellings for stylistic purposes. A parts-of-speech analysis reveals that the average tweet of the dataset contained 1.32 adjectives, 0.75 adverbs, 5.64 nouns and 2.38 verbs.

Table 1. Twitter data set

Construction of data set	Time interval	30 days (26 Sep-26 Oct 2014)
Characterization of data set	Number of tweets	1847
	Number of words	29753
	Words per tweet	16.1
	Emoticons	39
	Users	1312
	Users mentioned in text	1844
	Language	Norwegian
Manual annotation of data set	Negative tweets	410 (22.2%)
	Neutral tweets	1059 (57.3%)
	Positive tweets	378 (20.5%)

A group of three annotators were brought in for the manual annotation process. The annotators labeled each tweet as positive, negative or neutral, and each tweet was annotated by two people independently to ensure that the tweets' sentiments had been correctly understood and annotated. To calculate the reliability of the dataset, the joint probability of agreement and Cohen's Kappa were calculated. The equation for Cohen's Kappa is

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{1}$$

where p_0 is the relative observed agreement and p_e is the hypothetical probability of chance agreement when the observed data is used to calculate the probabilities of each annotator randomly saying each category. The overall agreement for the dataset was 83.9%, with a Kappa value of 0.73, which means that there is moderate and acceptable agreement among the annotators [16].

3 Related Work on Sentiment Analysis

A sentiment analysis system's objective is to extract the sentiment of a target, as defined above, on the basis of some textual resource. The task is normally handled as a natural language processing task at different levels of granularity, and we usually distinguish between unsupervised and supervised approaches [6]. Early work calculated sentiments at the document level [29] [23], though the focus has gradually shifted towards the sentence level [10] [15] and even the phrase level [30] [1].

The highest granularity level is the document level. At this level, one is concerned with determining the sentiment of each document as a whole [24]. For this level of granularity to be of any value, one usually wants to assume that each document expresses sentiments on a single topic. Corpora with documents such as customer reviews are very suitable for analysis at this granularity level. A more detailed level of granularity is the sentence level, where methods performing sentiment analysis attempts to determine the sentiment of single sentences. Finally, the finest level of granularity is at the entity level. In order to analyze the sentiments at the entity level one has to create a more holistic model that includes the target of the expressed sentiments. This, of course, requires more advanced linguistic computation and information modelling. Systems performing analysis at this level are very useful tools for performing structured sentiment summaries on entities, turning unstructured text into structured data.

With respect to sentiment analysis of news content the focus has been on longer texts, like online finance news [21] or product reviews [26] [29]. As noted in [20], the news domain is both less researched and understood. The authors experiment with sentiment classification within different domains attaining precision results between 75% and 95%. Their framework struggled with news article documents, yielding precision scores down to 75% due to difficulties in dealing with long and complex news documents.

Sentiment analysis is normally conducted following a two step process. First, you identify a text to be either objective or subjective. Subsequently, you take the subjective tweets and determine their polarity, i.e. assess whether they are negative or positive [23]. These two steps often make use of supervised learning methods. Supervised learners are often the methods of choice when annotated datasets are available. In the case of Twitter, there are means of obtaining datasets where the tweet classes can be determined automatically [22]. This enables the acquisition of large training datasets without the tediousness of manual annotation.

When using machine learning techniques for text classification, feature engineering is an important part of it. The features of a machine learning classifier are a selected subset of the measurable properties that define the documents in the corpus. Selection of the feature set is often performed as a combination of empirical selection by a domain expert and automated methods. The set of feature values for a given document is usually called the feature vector of the document.

In the following we will describe Naive Bayes, Support Vector Machines and Max Entropy in some more detail, as these machine learning techniques are used in our own experiment.

3.1 Naive Bayes

Naive Bayes(NB) is a fast and versatile classification algorithm that is widely used in supervised text classification systems, though it is often outperformed by more sophisticated classifiers like Support Vector Machines (SVM) [4]. The NB classifier is based on Bayes theorem, which specifies the relationship between the probabilities of two events A and B:

$$P(A|B) = \frac{P(A) \times P(B|A)}{P(B)} \tag{2}$$

In short, this theorem enables a classifier to calculate the posterior probability of B given A, using prior probabilities. In a NB classifier for tweets, the formula can be reformulated as follows:

$$P(c_p|\overrightarrow{d_j}) = \frac{P(c_p) \times P(\overrightarrow{d_j}|c_p)}{P(\overrightarrow{d_j})} \tag{3}$$

where $P(d_j)$ is the probability that a randomly selected tweet will be represented by d_j, and $P(c_p)$ is the probability that a randomly selected tweet belongs to class c_p. The classification function then is to find the class with the largest probability function given by the product of all the feature probabilities, given their class labels. This functionality is described by the equation below.

$$classify(f_1, ..., f_n) = argmaxp(C = c) \prod_{i=1}^{n} p(F_i = f_i|C = c) \tag{4}$$

The equation above shows the intuitive nature of Naive Bayes classifiers. In essence, we need to train our classifier by counting all the features and which classes they appear in, and use these frequencies to compute their probabilities. When classifying a tweet, we select the class which is given the highest product of the features given by the target feature vector.

3.2 Support Vector Machines

Support Vector Machines(SVM) is a relatively new technique for text classification and was first used for this purpose by Joachims in 1999 [14]. Compared to

the NB classifier, the SVM method is conceptually more complex and also more challenging to implement.

The central idea in SVM is to find the support vectors which maximize the space - the decision surface - between the two classes, i.e. finding the optimal separation between the features representing the two classes. The two support vectors are defined by the documents that lie closest to the decision surface.

The task of training an SVM classifier can be formulated as an optimization problem of finding the optimal hyperplane. Baeza-Yates & Ribeiro-Neto state this optimization problem as follows [2]:

Let H_w be a hyperplane that separates all documents in class c_a from all documents in c_b. Let ma be the distance of H_w to the closest document in class c_a and let m_b be the distance of H_w to the closes document in class c_b, such that $m_a + m_b = m$. The distance m is the margin of the SVM. The decision hyperplane H_w maximises the margin m.

When the optimized decision surface has been calculated, any future instance presented to the classifier is evaluated using their position in the space as represented by the features of this instance. The instance's position relative to the separation between the classes determines which class should be linked to the new instance.

3.3 Maximum Entropy

A Maximum Entropy(MaxEnt) classifier is a conditional probabilistic classifier. Implementations of it use logistic regression in order to find the probability distribution with the largest entropy, which - given by the Theory of Maximum Entropy [13] - should be the one best to represent the current state of knowledge, given precisely stated prior data [18].

Unlike the NB classifier, MaxEnt assumes no conditional independence for the features. This means that MaxEnt handles feature overlap better than the NB classifiers [7]. It also means that for text-only features, the MaxEnt classifier will often performs better given that most of the time we work with words that are conditionally dependent of each other.

[7] formulates the MaxEnt model in the following way:

$$P(c|d,\lambda) = \frac{exp[\sum_i \lambda_i f_i(c,d)]}{\sum_{c'} exp[\sum_i \lambda_i f_i(c,d)]} \tag{5}$$

where c is the class and d is the tweet. The numerical operations of the task of optimizing these lambdas are complex and and often lead to non-trivial and time-consuming implementations. For text classification tasks, MaxEnt classifiers have been shown to have an accuracy performance comparable to SVM [18].

3.4 Twitter Sentiment Analysis Approaches

Like in our work [7] use Naive Bayes, MaxEnt and Support Vector Machines for sentiment analysis of Twitter data. Sentiment data is acquired using a distant

learning approach. Positive (like :-)) and negative emoticons (like :-() at the end of tweets are interpreted as signs of positive and negative tweets. They experiment with Unigram and Bigram models in conjunction with parts-of-speech features. In their work the unigram model outperforms all other models, and SVM outperforms other classifiers.

A similar distant learning paradigm is adopted in the work of [22]. Their objective is to classify tweets as subjective versus objective. For subjective data they collect tweets ending with emoticons in the same manner as in [7]. Objective tweet data are obtained from crawling twitter accounts of popular newspapers like New York Times, Washington Posts, etc. As opposed to [7], [22] report that parts-of-speech and bigrams help improve the results.

Another interesting approach to sentiment analysis with Twitter is published by [3]. In their work polarity predictions from three websites are used as noisy labels to train a model, whereas 1000 manually labeled tweets are used for fine-tuning. An interesting aspect of their system is the use of syntactic features of tweets like retweet, hashtags, links, punctuation and exclamation marks in conjunction with features like prior polarity of words and parts-of-speech of words.

Our approach is in many ways similar to [7], though it has a wider scope and also addresses the issues of entities and temporal development. As opposed to their work, though, we also make use of semantic representations and find a slight improvement of accuracy when these semantic features are included. Also, just like in information retrieval [8] [27], we make use of linguistic techniques to gradually add more semantics into the whole analysis process.

4 Sentiment Analysis Approach

Given a news tweet x, the task of our sentiment analysis system is to determine whether x expresses a positive, negative, or neutral opinion. The system is split into two separate stages, subjective classification and sentiment classification:

- The subjective classification component first decides if the tweet contains a sentiment or not. If there is a sentiment represented, we call the tweet subjective and it is sent to the sentiment classification component for further analysis. Otherwise the analysis is terminated and the tweet is labeled neutral.
- The sentiment classification component's task is to categorize subjective tweets as either positive or negative.

The overall classification process is illustrated in Figure 2. Three machine learning techniques are used and evaluated for both classifiers. The tweets themselves are represented as sets of features that refer to both word properties and sentence properties of the tweets. Three different feature sets are tested for each machine learning techniques, giving us a total of nine runs for the classifiers.

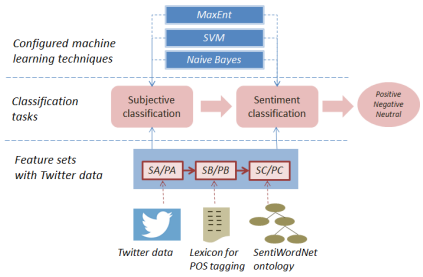

Fig. 2. The sentiment analysis process.

4.1 Configuration of Machine Learning Techniques

The machine learning techniques were extensively tested with different parameter values before the actual evaluation was carried out. Figure 3 lists the relevant parameters with the associated value ranges for each parameter. Two of the parameters were used for text vectorization: the range of N-grams used as features, and the Max document frequency for using the N-grams as features. Three parameters were for TF-IDF vectorizing: Use IDF, Smooth IDF, and Sublinear TF, all three of them boolean values. Finally, four algorithm-specific parameters were used: The Alpha parameter of the NB classifier, which is the Laplace/Lidstone smoothing weight, the C parameter in the SVM, which influences the margin of the SVM hyperplane, and lastly the MaxEnt-specific parameters C and penalty.

Parameter for machine learning technique	*Values*
N-gram range:	1-1, 1-2, 1-3, 2-2, 3-3
Use IDF:	True, False
Smooth IDF:	True, False
Sublinear TF:	True, False
Max DF:	0.5, 0.7, 0.9, 1.0
Alpha (NB-specific):	0.1, 0.3, 0.5, 0.7, 0.8, 1.0
C (SVM-specific):	0.1, 0.3, 0.5, 0.7, 0.8, 1.0
C (MaxEnt-specific):	0.1, 0.3, 0.5, 0.7, 0.8, 1.0
Penalty (MaxEnt-specific):	l1, l2

Fig. 3. Parameter combinations for optimizing the machine learning techniques.

The best parameter sets for the two classification tasks and the three machine learning techniques are shown in Figure 4. Even though there are few differences for the two classification tasks, the small deviations are important to the final outcome of the classifiers. Consequently, the rest of the experiment was conducted using the parameter values from Figure 4.

4.2 Syntactic and Semantic Enrichment of Feature Sets

For each of the classification tasks, three different feature sets were defined and evaluated. The simplest feature sets, SA for subjective classification and PA for

	Subjective classification			Sentiment classification		
	NB	SVM	MaxEnt	NB	SVM	MaxEnt
N-gram range	1-1	1-3	1-2	1-1	1-1	1-1
Use IDF	True	True	True	True	True	True
Smooth IDF	True	True	True	True	True	True
Sublinear TF	False	True	True	True	True	True
Max DF	0.5	0.5	0.5	0.5	0.5	0.5
Alpha	0.3			0.3		
C		0.7			0.7	
C			1.0			0.8
Penalty			l1			l2

Fig. 4. Paramter values for subjective classification and sentiment classification.

sentiment classification, contain weighted representations of the word tokens of the tweets, without any additional features at the sentence level.

A POS tagger from the TypeCraft project was used to tag the tweets with parts-of-speech information. This gives us additional information about syntactic and morphological properties of the tweets, e.g. the number of adjectives in a sentence or the use of negations. Feature sets SB and PB were enriched with such syntactic sentence level features, as earlier analyses suggested that there is some correlation between frequency of parts-of-speeches and tweet polarity.

Feature set SA
Word Features
Word tokens
Feature set SB
Word Features
Word tokens, POS-tag occurrences
Sentence Features
Number of exclamation marks, number of emoticons, number of adjectives in sentence, number of adverbs in sentence(except "ikke"), pronoun in sentence(binary), negation in sentence(binary)
Feature set SC
Word Features
Word tokens, POS-tag occurrences, subjectivity scores
Sentence Features
Exclamation marks, emoticons, adjectives in sentence, adverbs in sentence(except "ikke"), pronoun in sentence(binary), negation in sentence(binary), total subjectivity score from word polarities, total objectivity score, number of subjective words, number of objective words

Fig. 5. Feature set for subjective classification.

An important part of the experiment was to assess the value of combining standard classifiers with semantically enriched feature sets. The general idea was to associate identified entities in tweets with concepts (synsets) in WordNet. As WordNet does not exist for Norwegian, this extraction of concepts involved using Bing to translate Norwegian entities into their English counterparts. Having identified the relevant concepts, we used SentiWordNet to retrieve standard sentiments of the concepts.

SentiWordNet is an open sentiment lexicon, in which each synset of Wordnet is associated to three numerical scores Obj(s), Pos(s) and Neg(s), describing how objective, positive and negative the synset terms are [5]. The synsets, which may be considered concepts of a domain-independent ontology, are hierarchically organized and linked to sets of terms that are used to refer to the synsets in texts. A particular term may be part of several synsets if it can denote different things in different contexts. As an example, take the WordNet synset Good#1, which represents one of many interpretations of the term good. In SentiWordNet the

Feature set PA
Word Features
Word token
Feature set PB
Word Features
Word token, POS-tag occurrences
Sentence Features
Number of happy emoticons, number of sad emoticons, message length
Feature set PC
Word Features
Word tokens, POS-tag occurences, polarities
Sentence Features
Number of happy emoticons, number of sad emoticons, message length, total polarity score, number of positive words, number of negative words

Fig. 6. Feature set for sentiment classification.

Good#1 synset has an objective score of 0.25, a positive score of 0.75 and a negative score of 0.00. Adding together such scores for all identified concepts of a tweet, we get aggregated scores of the tweet's subjectivity and polarity.

The semantically enriched feature sets SC and PC include features that reflect the generation of aggregated sentiment scores from SentiWordNet concepts. The exact features used in the six sets are listed in Figure 5 and 6.

5 Evaluation

The annotated dataset contained a total of 1847 tweets about Norwegian news entities. There were 9 experimental runs with 10-fold cross validation for each classification task, one for each combination of machine learning technique and dataset. In addition we tested the two classification tasks on datasets of different sizes to verify their dependence on large-scale training data. Calculating the quality of the sentiments we used the notions of Accuracy and F1 with the following formulas:

$$Accuracy = \frac{T_p + T_n}{T_p + F_p + T_n + F_n} \tag{6}$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{7}$$

where Tp is the number of true positives (actual positive tweet estimated to be positive by classifier), T_n is the number of true negatives, F_p is the number of false positives, and F_n is the number of false negatives. The results of the subjective classification component are summed up in Figure 7. As we can see, the F1 values are rather similar across machine learning techniques and datasets for the subjective classification tasks. The Support Vector Machine approach has slightly better results than the other two techniques, with an average F1 score of 0.66 compared to 0.58 for Naive Bayes and 0.61 for Max Entropy. The performance of the sentiment classification task is somewhat better than for subjective classification, as indicated in Figure 8. Again the SVM approach has the highest scores, but all three techniques have F1 scores above 0.7 when SentiWordNet has been used to enrich the feature sets.

Subjectivity	Support Vector Machines		Naive Bayes		Max Entropy	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
Dataset SA	0.68	0.67	0.68	0.59	0.67	0.59
Dataset SB	0.72	0.67	0.69	0.58	0.70	0.63
Dataset SC	0.70	0.65	0.68	0.57	0.69	0.62

Fig. 7. Subjective classification with three machine learning techniques and three feature sets.

Sentiment	Support Vector Machines		Naive Bayes		Max Entropy	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
Dataset PA	0.74	0.70	0.74	0.69	0.72	0.67
Dataset PB	0.77	0.75	0.76	0.71	0.75	0.73
Dataset PC	0.79	0.78	0.75	0.72	0.77	0.75

Fig. 8. Sentiment classification with three machine learning techniques and three feature sets.

It is interesting to analyze the contribution of semantics in some more detail. Whereas the smallest feature sets, SA and PA, only consist of word tokens, the semantically enriched datasets of SC and PC contain information that is calculated from looking up concepts in SentiWordNet. If we compare the results from SA/PA with SC/PC, we can estimate the contribution of semantics in our sentiment analysis system. Figure 9 tells us the changes of F1 values when the simplest dataset is replaced with the semantically enriched dataset, while keeping everything else unchanged. Surprisingly, the addition of semantic features has a negative contribution on SVM and Naive Bayes in the subjective classification task. For the sentiment classification task the use of SentiWordNet improves the F1 scores for SVM by 11.4% and for Max Entropy by 11.9%. The improvement is less for Naive Bayes, though all techniques display a significantly higher F1 score when semantic features are introduced.

Change of F1 values	SVM	Naive Bayes	Max Entropy
Subjective classification with feature set extended with SentiWordNet sentiments (SA -> SC)	-3.0%	-3.4%	+5.1%
Sentiment classification with feature set extended with SentiWordNet sentiments (PA -> PC)	+11.4%	+4.35%	+11.9%

Fig. 9. Effect of including sentiments from SentiWordNet ontology.

It is also worth noting how the F1 scores improve with the size of the datasets. Figures 10, 11 and 12 show the F1 scores as a function of dataset size for Naive Bayes, SVM and Max Entropy. The scores improve very fast as the dataset is still below 350-400 tweets. With a dataset size between 400 and 1847 tweets the scores improve at a rather slow but steady pace, though they do not seem to have reached their maximum level when the full dataset is employed. Probably we would get even higher scores for all three machine learning techniques if we had a larger dataset available.

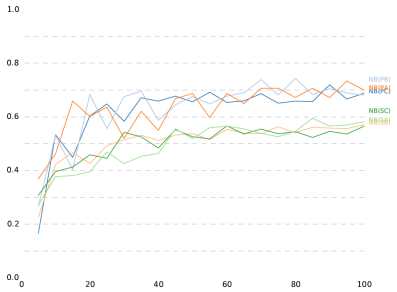

Fig. 10. F1 scores for Naive Bayes as a function of dataset size.

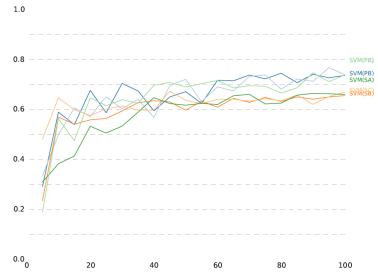

Fig. 11. F1 scores for Support Vector Machines as a function of dataset size.

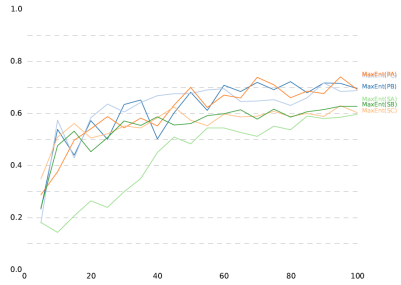

Fig. 12. F1 scores for Max Entropy as a function of dataset size.

6 News Entity Sentiments over Time

Estimating the sentiments of news tweets in general is not necessarily a very useful news service. The tweets span all kinds of topics, and the aggregated sentiment values combine the sentiments of events that probably have very little to do with each other. The sentiments may possibly reveal something about people's general attitudes or outlook, but sentiments make more sense when they are attributed to a particular target or aspect of this target. Moreover, the absolute sentiment values are of limited value to outsiders, as it is difficult to fully comprehend what a numerical value of sentiment actually mean. Twitter sentiment analysis becomes more useful when it is done at the level of entities for comparative analyses or trend analyses. Below we have extracted the news tweets about Erna Solberg, the Norwegian prime minister, from our initial data set. The subset about Erne Solberg consists of 662 tweets with a total of 10,974 tokens. A manual inspection shows that 307 tweets (46.4%) were neutral, 110 tweets (16.6%) were positive and 245 tweets (37.0%) were negative.

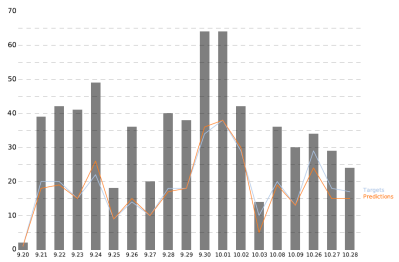

Fig. 13. Subjective tweets about Erna Solberg compared to tweets in total about Solberg.

Figure 13 shows the aggregated subjectivity for Erna Solberg tweets in a period spanning 19 non-consecutive days from 20th of September to 28th of October 2014. The red line is our estimated SVM sentiments for Erna Solberg over a period of little more than a month. The real sentiments, as indicated by the annotators, are given by the blue Target line. The grey bars in the background show the actual frequencies of tweets mentioning Erna Solberg. Interestingly, the share of subjective tweets to total number of tweets is fairly constant over time, though there are particular periods with a substantially higher share of subjective tweets.

Figure 14 shows the polarity differences from the Erna Solberg dataset during the same time period. The red Predictions line shows the aggregated differences between all positive tweets and all negative tweets per day. When the line is above the grey dotted Neutral line, it means that there are more positive than negative tweets on that particular day. The grey bars in the background show the total number of subjective tweets about Erna Solberg from day to day.

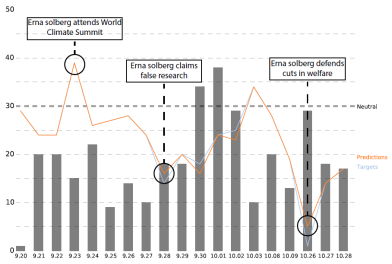

Fig. 14. Sentiments of Erna Solberg over time.

Interestingly, the polarity differences do not follow the pattern of subjective tweets per day. Clearly, there are periods in which people are strongly negative or strongly positive to the actions or statements of the Norwegian prime minister. The sentiments may change dramatically over just a few days, even though the number of subjective tweets does not change much. A closer look at the news in this time interval reveals some suspicious correlations. Around 21-22 September the positive sentiments of Erna Solberg come at a time when it is announced that she will take part in the World Climate Summit. Similarly, the opinions of the prime minister turn sour when it is revealed that the government referred wrongly to some research results, and even more when Solberg had to go public and defend cuts in well-fare.

7 Conclusions

This paper describes a Twitter sentiment analysis component that is developed as part of NTNU's SmartMedia program. Three different classifiers, SVM, Naive Bayes and Max Entropy, have been implemented and evaluated as part of this work. The component has been tested on a manually annotated Norwegian news dataset from Twitter. Additional features from lexical resources and sentiment ontologies have been included to examine the contribution of deeper syntactic or semantic analyses of text.

The results suggest that the three approaches are not very different in terms of precision and dependence on data set size, but the choice of feature set is important. In total SVM had the highest precision of the three in sentiment classification and was substantially better for very small data sets or poor feature sets. Maximum Entropy was efficient in subjectivity classification when more informative feature sets were available.

Adding semantic features with the help of SentiWordNet leads to a substantial improvement of the sentiment classifier, but not of the subjective classifier. Both SVM and Max Entropy see an improvement of more than 11% when SentiWordNet is consulted to enrich the feature sets for the sentiment classifier. It is difficult to assess why we do not see a similar improvement in subjective classification, though.

We also analyzed to what extent variations in sentiments coincide with important events dealing with these entities. The analysis reveals that sudden changes of sentiments can usually be attributed to concrete news events that are heavily reported in the media. There is, however, little correlation between the share of subjective tweets and particular news event. It seems that people do not get more emotional when major events take place, but the polarity of their emotional tweets seem to correlate well with their opinions of the underlying news.

References

1. A. Agarwal, F. Biadsy, and K. R. Mckeown. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 24–32. Association for Computational Linguistics, 2009.
2. R. Baeza-Yates, B. Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.
3. L. Barbosa and J. Feng. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 36–44. Association for Computational Linguistics, 2010.
4. A. Bermingham and A. F. Smeaton. Classifying sentiment in microblogs: is brevity an advantage? In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1833–1836. ACM, 2010.
5. A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422. Citeseer, 2006.
6. R. Feldman. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89, 2013.
7. A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1:12, 2009.
8. J. A. Gulla, P. G. Auran, and K. M. Risvik. Linguistics in large-scale web search. In *Natural Language Processing and Information Systems*, pages 218–222. Springer, 2002.
9. J. A. Gulla, A. D. Fidjestøl, X. Su, and H. Castejon. Implicit user profiling in news recommender systems. 2014.
10. M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.
11. M. Hu, S. Liu, F. Wei, Y. Wu, J. Stasko, and K.-L. Ma. Breaking news on twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2751–2754. ACM, 2012.
12. J. E. Ingvaldsen, J. A. Gulla, and Ö. Özgöbek. User controlled news recommendations. In *Proceedings of the Joint Workshop on Interfaces and Human Decision Making for Recommender Systems co-located with ACM Conference on Recommender Systems (RecSys 2015)*, 2015.
13. E. T. Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.
14. T. Joachims. Making large scale svm learning practical. Technical report, Universität Dortmund, 1999.

15. S.-M. Kim and E. Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1367. Association for Computational Linguistics, 2004.
16. J. R. Landis and G. G. Koch. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, pages 363–374, 1977.
17. B. Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.
18. C. Manning. Maxent models and discriminative estimation. *CS 224N lecture notes, Spring*, 2005.
19. M. Mendoza, B. Poblete, and C. Castillo. Twitter under crisis: Can we trust what we rt? In *Proceedings of the first workshop on social media analytics*, pages 71–79. ACM, 2010.
20. T. Nasukawa and J. Yi. Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture*, pages 70–77. ACM, 2003.
21. P. C. S. Njølstad, L. S. Høysaeter, W. Wei, and J. A. Gulla. Evaluating feature sets and classifiers for sentiment analysis of financial news. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on*, volume 2, pages 71–78. IEEE, 2014.
22. A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326, 2010.
23. B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics, 2004.
24. B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.
25. T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.
26. F. Simančík and M. Lee. A ccg-based system for valence shifting for sentiment analysis. *Research in Computing Science*, 41:99–108, 2009.
27. G. Solskinnsbakk and J. A. Gulla. Combining ontological profiles with context in information retrieval. *Data & Knowledge Engineering*, 69(3):251–260, 2010.
28. M. Tavakolifard, J. A. Gulla, K. C. Almeroth, J. E. Ingvaldesn, G. Nygreen, and E. Berg. Tailored news in the palm of your hand: a multi-perspective transparent approach to news recommendation. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 305–308. International World Wide Web Conferences Steering Committee, 2013.
29. P. D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics, 2002.
30. T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics, 2005.