OQAM

Let the sentiment be your friend

A short introduction to NLP, BERT and news sentiment in investing

Kristoffer Nordström & Thorbjörn Wallentin

Intro

What does one of the characters from the Muppets Show have to do with investing and sentiment?

We at OQAM, a quantitative asset management firm, are fascinated by the latest technological developments. Recent advances in the field of natural language processing (NLP) have led to new ways of systematically analysing vast amounts of text data and hence created new ways of generating actionable signals to the investment process.

In this whitepaper, we explore how to use NLP in the investment process. We are primarily targeting people with an interest in finance and who want to understand/get an update on the recent developments within natural language processing. We start by introducing the subject. From there we move on to show how to build a sentiment index by analysing press releases. Lastly, we finish by presenting a quantitative investment strategy based on the sentiment index.

The strategy we create manages to avoid the pandemic shock in March 2020 and increase the risk-adjusted return compared to the benchmark index. Thus, showing promising results for using automatically analysed text data as an input in the investment process.

What is NLP?

NLP, or natural language processing, is a field within computer science concerned with the task of giving computers the ability to understand human written language. This is mainly achieved with the help of artificial intelligence and statistical methods.

Applications leveraging NLP are encountered frequently in our daily lives; when using a website's search bar, virtual assistants (Siri/Alexa), spam filters, chatbots, and Google Translate, to mention a few examples. In business, there are many different applications, many of which relates to automation.

For NLP to work, language needs to be encoded in a set of rules that a computer can understand and follow. This is a complex task as the meaning of a word is dependent on the context, like in the following examples, where the same words have a completely different meaning:

Let's **eat**, **grandma**. Let's **eat grandma**.

To correctly handle this, modern NLP methodologies uses deep learning to model the context of the sentence.

In recent years the field has seen a high activity of research and lots of advancements, one breakthrough was Google's release of a new text-based model called BERT in 2018, significantly improving the capacity of past models to correctly understand context. This is the model we have been using at OQAM and in this whitepaper. Last month OpenAI opened access to an NLP model called GPT-3, giving developers and researchers access to one of the most prominent language models created, showing that the field is still evolving at a fast pace.

NLP at OQAM

We at OQAM decided to initiate our first NLP research project in the summer of 2020. The project was initiated as a summer project by two interns. Given that most of the work within NLP is conducted on the English language, and thus quite thoroughly researched, we decided to focus on the Swedish language.

Our ambition was to investigate if we could predict the category and sentiment of stock-specific press releases by using NLP. We believe that NLP is an important tool to incorporate alternative data sources such as news and press releases into the quantitative investment process.

Currently, we are evaluating a lab trading strategy based on NLP which has been implemented with a low risk allocation. The strategy has been live since mid-2021 and we are continuously monitoring its characteristics and evaluating potential next steps.

It all starts with the data

To successfully build NLP applications, it all starts with good data. A spam filter for instance gets better and better the more training data, consisting of correctly categorized e-mails (spam and non-spam e-mails), it can use.

In our research, we wanted to ensure we had the right training data for our task to interpret Swedish press releases. Out of our total data set of 60,000 press releases, we manually annotated 3,000 of these regarding sentiment and category. It was a tedious but central task as we believe the NLP model should be a digital extension of us, reflecting our views and opinions as investors.

To give a brief insight into the task of annotating the training dataset, consider the following press release:

> "AAK expects a higher operating profit for the third quarter than previously forecasted"

We decided that it should be categorised as an earnings-related press release and its corresponding sentiment should be positive.

What is BERT?

The acronym BERT stands *for Bidirectional Encoder Representations from Transformers* and reveals that the model is based on several recent advances within the field.

Compared to previous methods, BERT is a significant improvement to how AI models understand the context of text and words. For example, before BERT, NLP models such as Bag-of-Words often looked at the number of occurrences of positive and negative words to calculate sentiment. In some domains this may yield reasonable accuracy, however, within the financial domain, it struggles, as it does not capture the context around the words.

The example below shows the importance of contextual understanding when conducting sentiment analysis.
If the words “reduced” and “costs” yield negative scores and the word “increased” yield a positive score, the two sentences are falsely classified. BERT, however, sets words in context to each other, and hence understands that reduced costs are positive while increased costs are not.

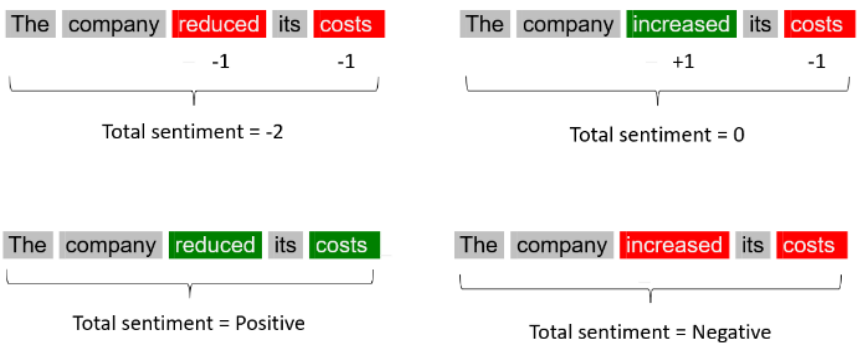
How to use BERT

Before a BERT model can be used to solve a text-related problem it must be trained in two steps. First, the model needs to be taught the language, for example Swedish. This step is called *pre-training*. Then the model needs to be taught the specific task, this is called *fine-tuning*.

The first step is accomplished by feeding the model massive amounts of text data. This process is computationally expensive and needs a lot of data, as the model analyses billions of text samples to build an understanding of the language. Luckily, the National Library of Sweden already had done the heavy work for us and published pre-trained models that we could use.

The second step is to train the model on a specific task. This is done similarly as above, but with samples from the specific task. In this whitepaper, we want to leverage a BERT model to classify press releases either as positive, neutral, or negative. Meaning that we need to show the model samples of press releases from each of the three classes.

> **The role of the National Library of Sweden and language models**
> The National Library of Sweden, currently have access to one of the largest collections of Swedish texts, making it is possible for them to pre-train language models on vast amounts of data. They publish their models online, enabling companies, organisations, and individuals to leverage language models that otherwise would be very expensive to develop.



Example showing the importance of contextual understanding in sentiment analysis

Constructing a sentiment index

In this section we will show how to build a sentiment index, a time series capturing and aggregating the sentiment of all press releases communicated by companies listed in Sweden.

We are using press releases from large- and mid-cap companies listed on the Swedish stock exchange. The universe is a snapshot of today, and some selection bias may therefore occur in the index. The total data set consists of roughly 60,000 unique press releases from 2010 and forward. The first step of creating a sentiment index is to get the individual sentiment for each press release.

Make BERT work for us

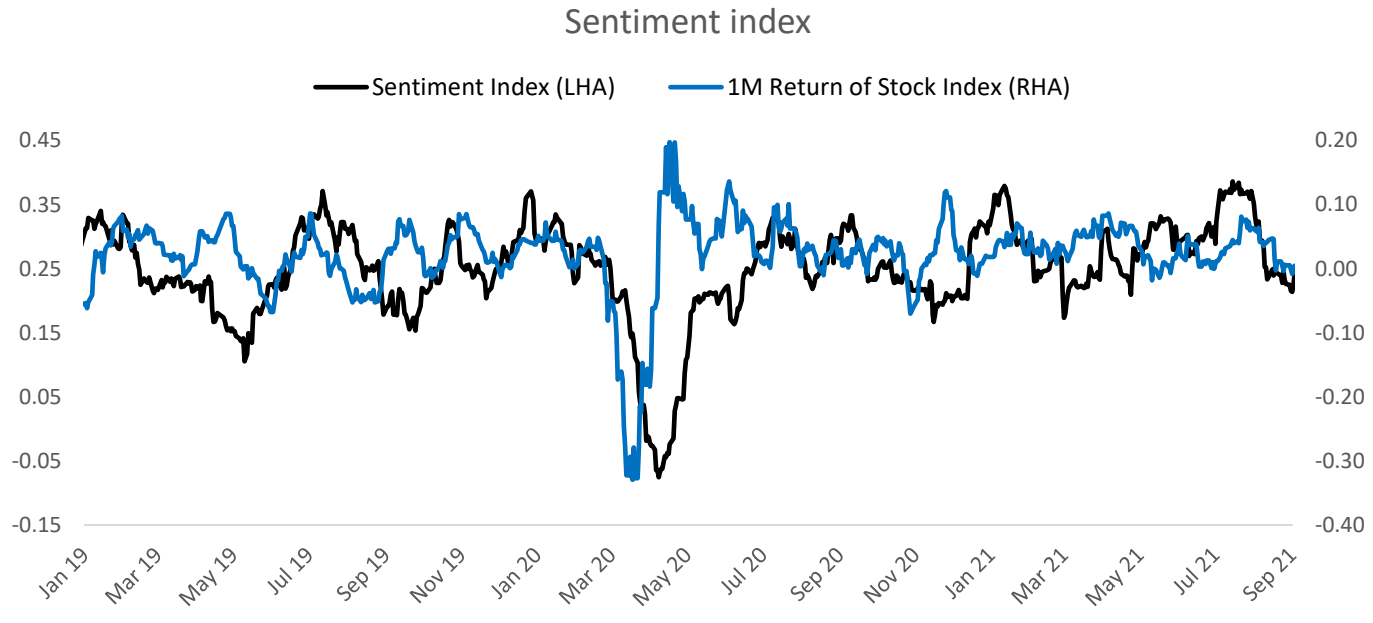
One could manually go through each press release and label it as either positive, negative, or neutral, and aggregate these into an index, but going through 60,000 samples takes a lot of time, it is also tedious if the index should be updated regularly. Hence, we want to go through a subset of all press releases, fine-tune a BERT model on these, and then let it do the work for us on the rest of the samples. This procedure makes it possible to classify new press releases efficiently without human supervision.

As mentioned above, we manually annotated 3,000 samples. We then fed these into the BERT model and let it classify the remaining 57,000 press releases for us.

The sentiment index

There are multiple ways of constructing an index, but we decided to keep it simple. For each day we count the respective number of positive, negative, and neutral press releases. We calculate the difference between positive and negative press releases and divide it by the total amount of press releases that day. Lastly, we take the 25-day (5 weeks) moving average and shift all data backward one day. This smooths out the time series and makes sure that any look-ahead bias is avoided. The resulting sentiment index can be seen in the figure below.

The sentiment index reaches its lowest value during the initial market reaction of the Covid-19 pandemic and increases during the summer, coherent with how the market moved during the same time. The sentiment index also seems to correlate with the 1 month return of the stock index.



Sentiment index from January 2019 to September 2021 compared to monthly returns in a broad Swedish stock index

Improving the risk-adjusted return with sentiment

With help of the sentiment index, we take it one step further and investigate if it's possible to use it to generate better risk-adjusted returns. We limit ourselves to one tradable asset, a broad benchmark index over the Swedish stock market. And the only two alternatives are either full exposure or no exposure.

By looking at the sentiment index compared to the monthly return of our stock index it is possible to spot some correlation between the returns and the sentiment. This correlation is most prominent during the pandemic shock in March 2020.

We continue the analysis by dividing trading days into three groups, one where the sentiment index has increased for the past month, one where it has stayed flat, and one where it has decreased. By doing this it is possible to investigate if there is any difference in the distribution of daily returns for each group respectively.

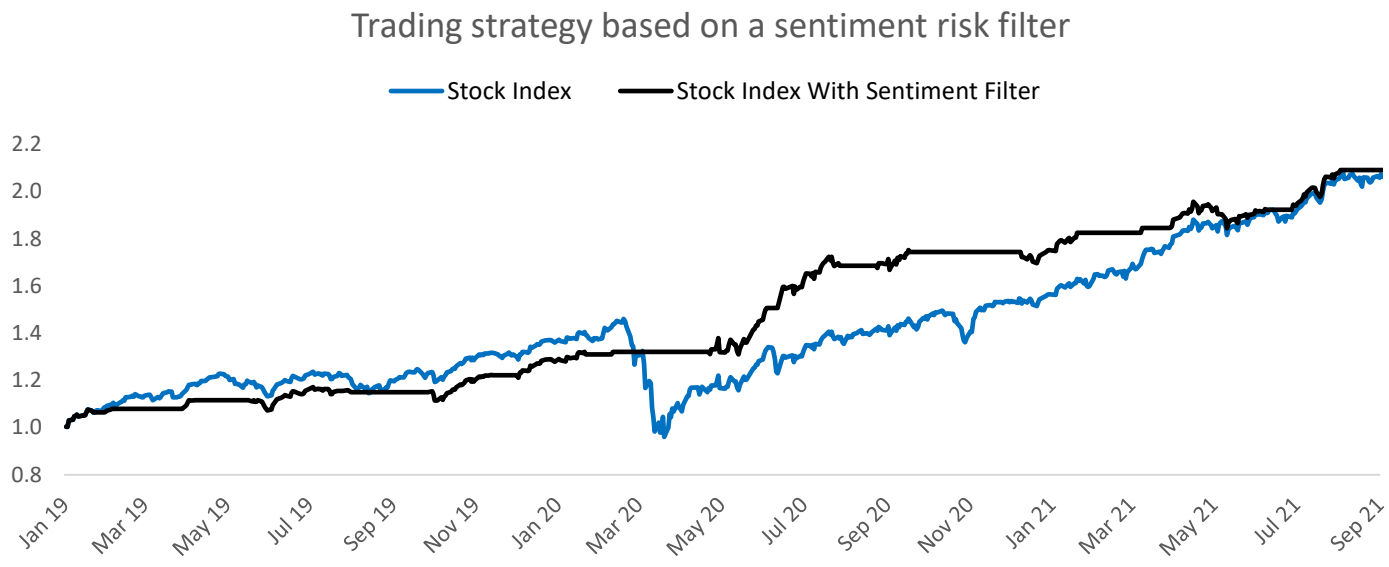
Sentiment	Average return	Std of returns	Group size	Return per std
Decrease	0.04%	1.61%	246	0.025
Flat	0.01%	1.05%	227	0.010
Increase	0.29%	0.93%	196	0.312

As seen in the table, there are more days where the sentiment is flat or decreasing. Further, days with rising sentiment have a higher average return while the standard deviation is smaller compared to days with a decreasing sentiment. The ratio between the mean return and the standard deviation is nearly 10 times greater for days with an increasing sentiment compared to days with a decreasing sentiment.

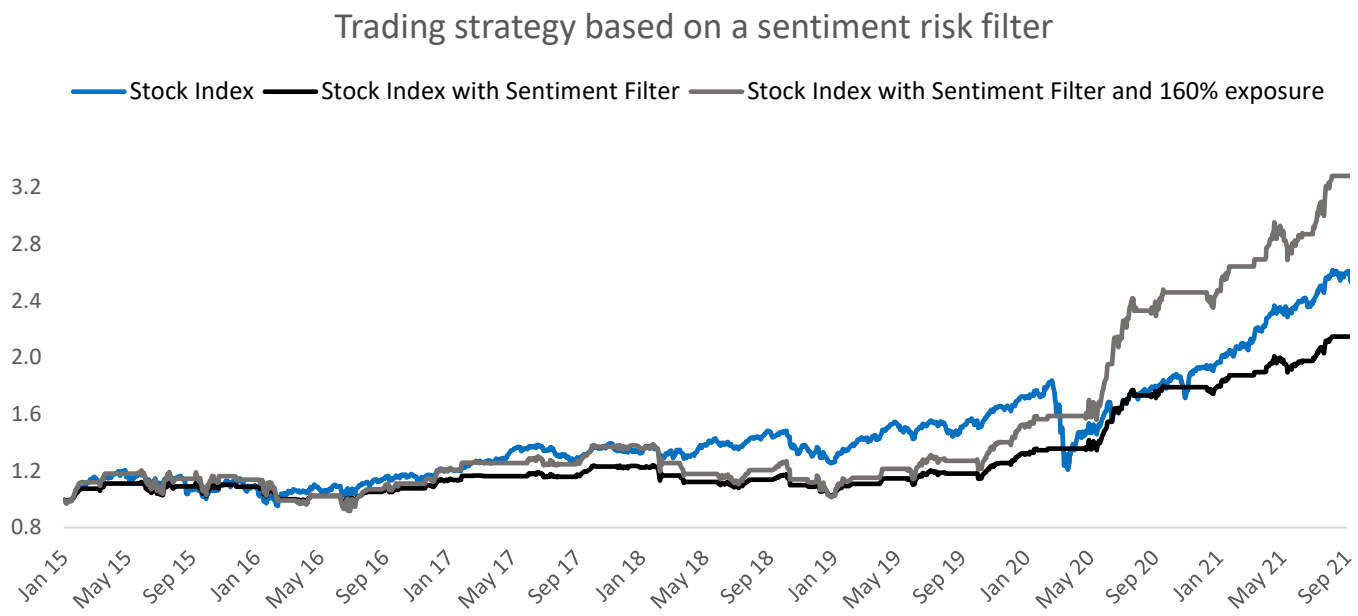
This analysis supports the thesis of a negative correlation between a rise in sentiment and risk in the market. Hence, we want to build a model with exposure to the stock index when the sentiment is rising and no exposure when the change in sentiment is flat or negative.

We implement this strategy by using a moving average approach. If the 25-day moving average sentiment value is above the 50-day moving average sentiment value, we have 100% exposure to the stock index, and otherwise no exposure at all. The result of the strategy is seen in the figure below.

The strategy successfully performs on par with the underlying index in terms of total return. Since the total time in market is lower for the strategy, it reduces the



Trading strategy based on sentiment. Stock index with and without sentiment filter



Trading strategy based on sentiment. Stock index with and without sentiment filter as well as increased exposure to even risk between strategies

overall market risk, while keeping approximately the same upside. The strategy successfully managed to stay out of the market during the drawdown in March 2020 and later put on exposure when the sentiment increased again.

Although the strategy shows some attractive characteristics during this period, to give the reader some perspective, the same strategy performs worse if we extend the time window back to 2015. As seen in the figure above, the strategy misses a lot of the performance in exchange for down-side protection. Until mid-2019 the strategy yields near-zero returns.

Because the stock index exhibits more variance, it is possible to adjust the exposure in the sentiment strategy to match the variance of the index. This yields an increase in exposure during on-signals from 100% to 160%. The sentiment strategy with increased exposure can be seen in the figure above. The increase in exposure gives a higher total return, while still limiting the downside.

Summary

The last couple of years' advancements in the field of NLP has opened new possibilities that hardly were imaginable in the past.

In this whitepaper, we have shown how to create a simple sentiment filter to increase the risk-adjusted returns of an index. Though the strategy needs more work before deployment, it gives a peek of what can be accomplished with the help of data and the latest NLP technologies. In general, interest in NLP within finance have been growing rapidly the last couple of years and potential use cases keep expanding. This article focused on a simple use case deploying a risk sentiment filter. In real life investors analyse sentiment in real-time, using all different sorts of data (Twitter/social media, news, economics, etc.) to help them manage their risk. Possible use cases could also be found for instance within the processes dealing with corporate actions. Making life hopefully smoother for asset managers focusing on large equity universes.

Read more

A good and illustrative guide on how BERT works:
https://jalammar.github.io/illustrated-bert/

The research team at the National Library of Sweden, a must if you want to stay up to date with cutting edge Swedish language models:
https://kb-labb.github.io/

Collection of state-of-art AI models, free to use:
https://huggingface.co/

Google's AI-blog, must follow for everyone interested in the field:
https://ai.googleblog.com/