Discovering the sentiment in finance's unstructured data

Introducing SentiMine

LSEG LABS

Contents

Authors

Kelvin Rocha
PhD, Lead Data Scientist
LSEG Labs Singapore

Joe Whyte
Innovation Lead
LSEG Labs Singapore

Kai Xin Thia
MS, Senior Data Scientist
LSEG Labs Singapore

Benjamin Chu
MS, Senior Data Scientist
LSEG Labs Singapore

Mihail Dungarov
MS, CFA
Senior Data Scientist
Refinitiv Workflow
Intelligence

1. The unstructured data explosion: why SentiMine matters

Today, it is estimated that at least 80% of the existing data is unstructured[1, 2]. This has led to an increased need for financial analytics companies around the world to use artificial intelligence (AI), machine learning (ML) and natural language processing (NLP) to parse textual data hundreds of thousands of times faster and more accurately than humans for financial analysis and insight generation.

In line with the increasing demand for AI/ML and NLP solutions, there have been regulatory changes in the market, for example, the separation of trading commissions paid by investment firms to brokers, compensation for research and greater transparency around trades. This change means there is greater demand than ever to show the value of unstructured data by making insights consumable from content like research and transcripts.

In response to the increased demand for more advanced and scalable ways of consuming unstructured content, LSEG Labs created SentiMine, a novel discoverability tool specialised in highly complex financial documents such as equity research reports (ERR) and transcripts. SentiMine automatically locates and assesses the sentiment of over 100, very specific financial-related themes, thought of by a selected panel of Refinitiv's equity experts as key drivers of equity performance. It is engineered to assist equity analysts in the discovery of trends and contrarian views that would otherwise be extremely difficult to do given the ever-increasing amount of information. SentiMine also helps to identify controversies in transcripts, where analysts and company representatives show very different points of view. With a performance, measured as weighted-F1 score, above 70% for all covered themes, SentiMine is a reliable tool that is primed to be essential to accelerate any equity research.

1 Amir Gandomi, Murtaza Haider, 'Beyond the hype: Big data concepts, methods, and analytics,' International Journal of Information Management, Volume 35, Issue 2, 2015, Pages 137-144.

2 Christine Taylor, 'Structured vs. Unstructured Data', www.datamation.com, 28 March 2018.

2. An introduction to SentiMine

Consuming and looking for drivers of equity performance in unstructured financial content such as equity research reports and transcripts, by searching, reading and analysing, is time-consuming and unscalable. It has also caused the underutilisation of these valuable sources of information.

Take the case of an equity research analyst who could be covering dozens of stocks. Hundreds of documents a day are available for use in analysis and thesis generation. These documents not only include equity research reports and transcripts, but also news, annual reports and filings among many others. They might contain thousands of words and numerous tables and figures, making this task of identifying potential drivers of equity performance a daunting one. This valuable data often going to waste is what motivated the team at LSEG Labs to create SentiMine.

SentiMine is an advanced discoverability tool for unstructured financial content that surfaces potential drivers of equity performance and sentiment from millions of text-heavy documents. SentiMine does so by breaking down the financial content on any equity into more than 100 relevant themes for which a continuous outlook (from positive to negative) is given, providing a high-level understanding of the equity and quickly identifying contrarian views.

A problem faced by every investment manager – equity analyst, portfolio manager or investment strategist – is keeping up to date with everything that is published about an equity they are covering. Figure 1 shows SentiMine's Thematic Overview for equity research reports. For any given equity, for a user-defined time frame, SentiMine identifies the corpus of research reports that were published and surfaces the key themes (from the most to the least frequent), and the analyst outlook for each of those themes that was mentioned across all the reports selected. A user can filter to only look at reports by specific analysts they care about, or by specific themes they deem most important for that equity. In minutes, a user can understand what key themes are being mentioned across all the reports, as well as the analysts' outlook for each theme, saving hours of work reading, analysing and synthesising.

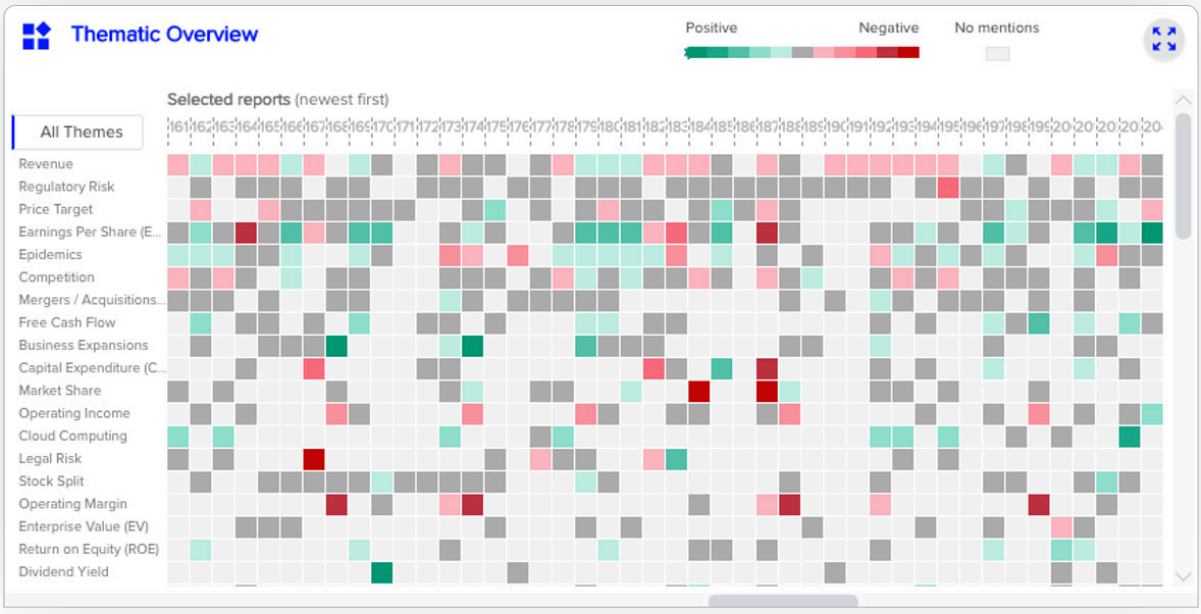


Figure 1 Thematic Overview for equity research reports (ERR) in SentiMine. Each column represents a research report. SentiMine themes are shown on the left, ranked by how frequently they appear on the reports. The more green the tiles are, the more confident the system is that the sentiment is positive. On the other hand, the more red they are, the more confident the system is that the sentiment is negative. Neutral sentiment is shown in grey.

3. How SentiMine works

Here, we walk through how SentiMine works and why this is significant. Section 3.1 outlines how themes were identified and how they are analysed within a financial document. Section 3.2 discusses how the SentiMine's theme-based sentiment models work and how scores are aggregated. Section 3.3 reviews the model training and annotation process. Section 3.4 is focused on how SentiMine works for transcripts and Section 4 highlights business use cases. Finally, in Section 5, we conclude with suggestions for future research work.

3.1 Theme identification and search model

LSEG Labs researched and identified more than 100 relevant themes that are considered by Refinitiv's subject-matter experts (past equity analysts, traders, etc... with many years of experience) to be potential drivers of equity performance. These themes have been classified into five core, homogeneous groups, based on their nature:

1. Accounting
2. Valuation
3. Business
4. Risk
5. Company Management

Figure 2 shows some examples for each of the groups.

The Labs team also identified themes that are very specific to the consumer retail, financial, telecommunications and technology sectors, and that fall into one of the five groups above. For example, 'Cloud Computing' is a business driver that is key for the technology sector.

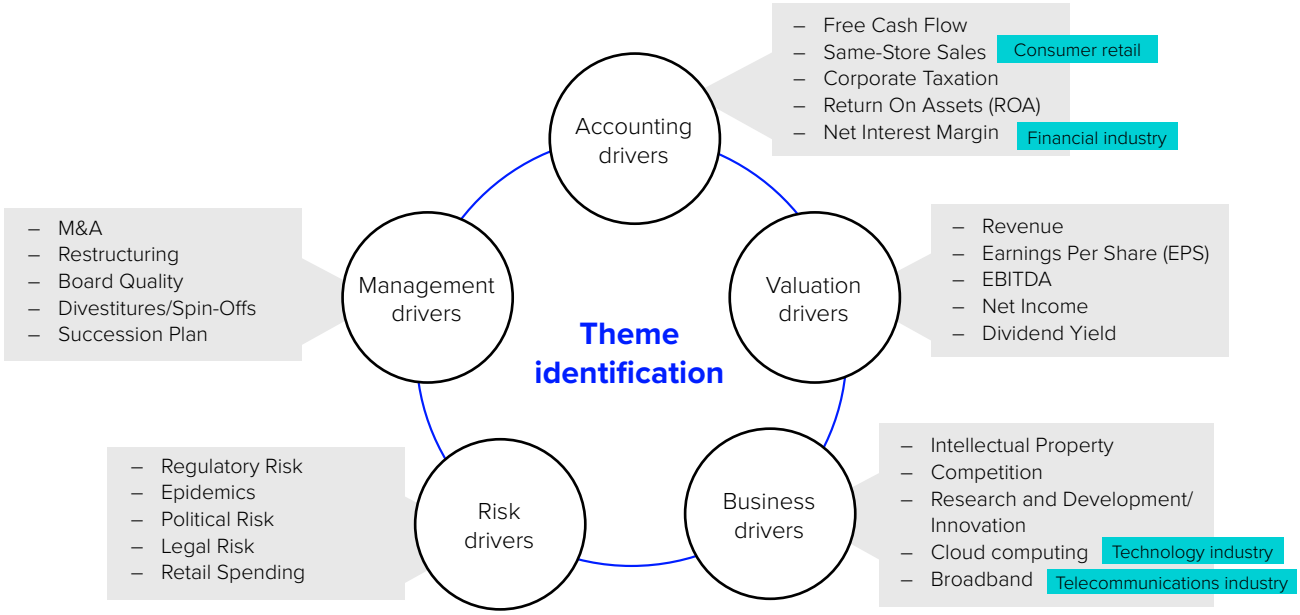


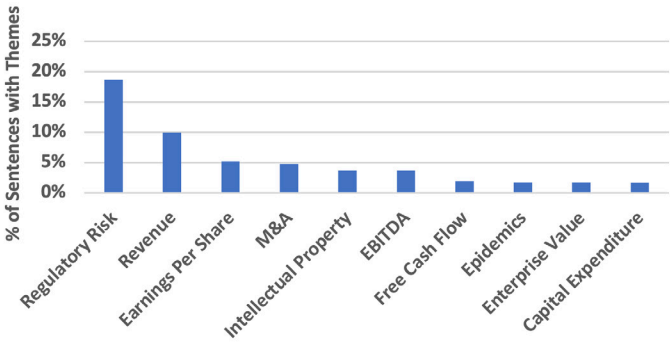
Figure 2 SentiMine is comprised of five homogeneous groups, based on their nature: Accounting, Valuation, Business, Risk and Company Management drivers. Some themes are very specific to sectors such as telecommunications, technology, financial or consumer retail. This figure shows some of the more than 100 themes in SentiMine.

Similarly, 'Net Interest Margin' is an accounting driver very specific for the financial sector, and 'Broadband' is a business driver for the telecommunications sector. These specialised themes are thought of as drivers of equity performance in their respective industries and having them in SentiMine is important for equity analysts covering those sectors. In the future, we might add themes related to other business sectors.

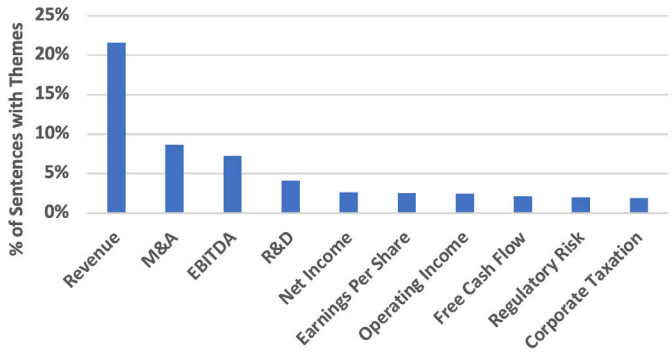
Once themes were identified, a search model was built into SentiMine. This model was carefully configured for each of the themes in such a way that it understands the nuances of the financial language and can precisely capture theme occurrences in the text corpus, while removing noise.

Figure 3 shows some of the most frequent themes in both equity research reports and transcripts, as a percentage of sentences containing any of the themes. As readers can see, the theme distributions are not exactly the same, with 'Revenue' being mentioned in more than 20% of the sentences with themes in transcripts, but less than 10% of the sentences with themes in equity research reports. On the other hand, 'Regulatory Risk' appears in 18% of the sentences with themes in equity research reports, but only in 2% of the sentences containing themes in transcripts. The reason for these differences is that equity research reports and transcripts are different types of financial documents with different characteristics and as such, SentiMine treats them differently. For example, in equity research reports, a single analyst spends more time explaining his/her ideas on the company he/she is writing the research on and tends to use a very technical language usually backed up by financial numbers. On the other hand, texts in transcripts tends to be more colloquial, as it is the result of a conference call between the company representatives and various analysts.

Top 10 themes in equity research reports



Top 10 themes in transcripts



Figures 3A and 3B Distribution of top 10 themes in equity research reports (ERR) and transcripts. Notice how the distributions differ. In fact, they are different types of financial documents with different characteristics. SentiMine's theme-based sentiment models treat them differently.

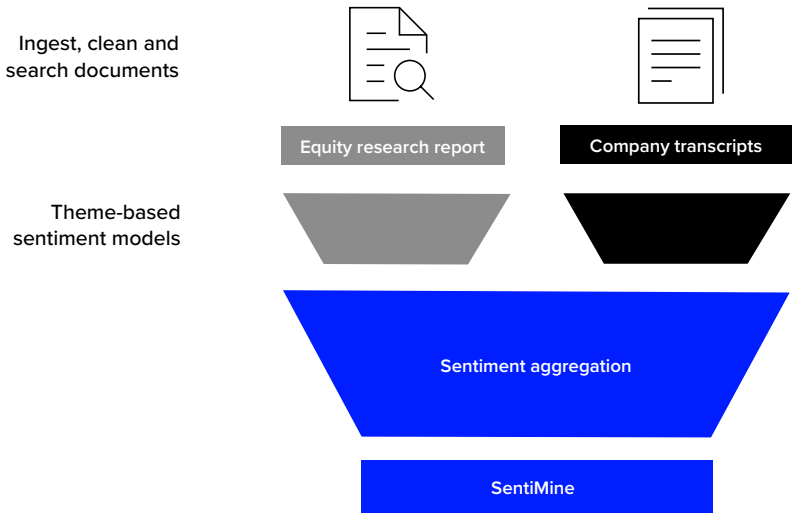


Figure 4 SentiMine's workflow. Documents are ingested, cleaned and searched before assigning sentiment to the different themes found.

3.2 Sentiment workflow and model description

Figure 4 shows SentiMine's workflow. When a document is passed through SentiMine, the first stage of the process is to prune it so that disclaimers and other related components are removed from the document, as it does not make sense to extract sentiment from them. Similarly, text and captions from tables and figures are removed, as their insights are usually addressed on the document corpus.

In the next stage, the theme search model is used to identify the text fragments where themes appear in the text. SentiMine then uses state-of-the-art sentiment models that have been carefully fine-tuned for these themes. Each of these models contains a neural network that has been trained to learn and understand contextual relations between the themes in the text, the theme's surroundings and the corresponding sentiment or outlook.

For every text fragment containing a theme, the corresponding theme-based model inside SentiMine provides the probabilities of that text fragment being positive, negative or neutral with respect to the given theme. These three probabilities are then converted into a single score that oscillates between -100 to 100. If the score is 100 or very close to it, there is a high chance the sentiment is positive. On the other hand, if the score is -100 or very close to it, then there is a high chance the sentiment is negative. Scores around 0 mean the sentiment is neutral (i.e., neither positive nor negative).

For example, the sentence: 'The company has benefitted from a first mover advantage in the cloud market and has sustained revenue growth at scale over the last two years' would be scored very positively for both 'Cloud Computing' and 'Revenue'. This because the probability of being positive with respect to these two themes is very large, as compared to being neutral or negative.

In the final stage of the process, the themes' sentiment scores are averaged at the document level, creating a single sentiment score per theme and per document, between -100 (highly likely to be negative) to 100 (highly likely to be positive), where 0 implies neutrality. This can be seen in SentiMine's 'Thematic Overview' (Figure 1), where each column represents a document and each coloured tile (green, grey and red) represents the sentiment (positive, neutral and negative) of each theme.

The LSEG Labs theme search and theme-based sentiment models are at the heart of SentiMine and, as described in the next section, together they have proven to deliver high-performance results when predicting sentiment across our selected themes.

3.3 Model training and performance

We have trained and tested SentiMine using hundreds of thousands of text fragments containing our key themes. These text fragments were randomly extracted from more than 3 million equity research reports (dated 2016 to 2020) and from more than 300,000 earnings call transcripts (dated 2000 to 2020), both covering global equities.

To ensure high-quality data, on one of the processes, each text fragment was manually labelled against a theme by three external individuals from a pool of dedicated taggers. Before and throughout the labelling process, each of these individuals was continuously tested via hidden test questions to make sure they were doing a high-quality job. If their test scores fell below a certain threshold, then all their annotated labels were excluded from the data. On another process, a subset of text fragments containing more difficult themes was carefully annotated by a team of internal, financial NLP experts.

SentiMine's theme-based sentiment models then utilise this labelled data in a cross-validation training process to match the best-performing models with their corresponding hyper-parameters. Model performance is always measured on unseen test data and uses a weighted-F1 score generating a positive, neutral and negative classification.

The intensive labelling and training process combined with the large amount of labelled data used and the high level of model sophistication, allow SentiMine to achieve a high weighted-F1 score (above 70%) on all covered themes. Table 1 shows the weighted-F1 performance statistics for the different driver groups. As shown, all driver groups have an average performance of 70% or above, with some performing at an above 80% confidence.

Driver group	Average	Minimum	Maximum	Standard deviation
Accounting	73.4%	70%	80%	3.70%
Business	73.0%	70%	84%	3.06%
Company management	75.2%	70%	84%	6.42%
Risk	74.4%	70%	80%	3.13%
Valuation	73.0%	70%	80%	2.77%

Table 1 Weighted-F1 performance statistics for the different driver groups in SentiMine. The weighted-F1 score uses positive, neutral and negative sentiment as three different classes.

3.4 SentiMine for transcripts

While equity research reports are written by analysts who provide their comments and recommendations on whether to buy, hold or sell shares of a company, transcripts are generated by the company itself during call presentations. Figure 5 shows the different types of transcripts covered by SentiMine. The majority of transcripts (70%) are derived from earnings conference calls held on a quarterly basis.

During these earnings calls, a presentation given by the company is followed by a Q&A session where analysts ask company representatives specific questions. Since the tone of presentation is often different than that of the Q&A (it is generally more positive), SentiMine treats the two call sections differently by applying a different set of theme-based sentiment models. Furthermore, to account for the varying sentiment added by analysts during the Q&A section, SentiMine further splits the Q&A session into questions and answers to give a more nuanced picture of the company and the analysts' opinion of it. Figure 6 shows how the different transcript sections are split in SentiMine in the 'Thematic Overview' pane.

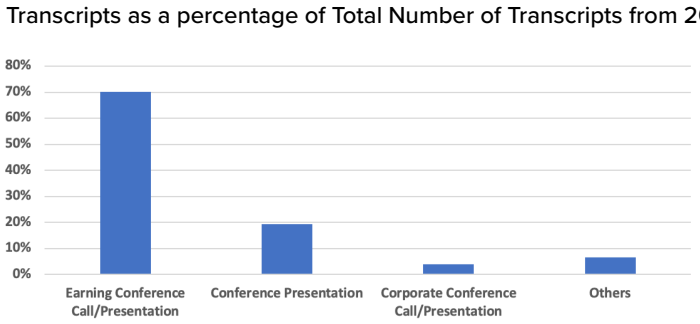


Figure 5

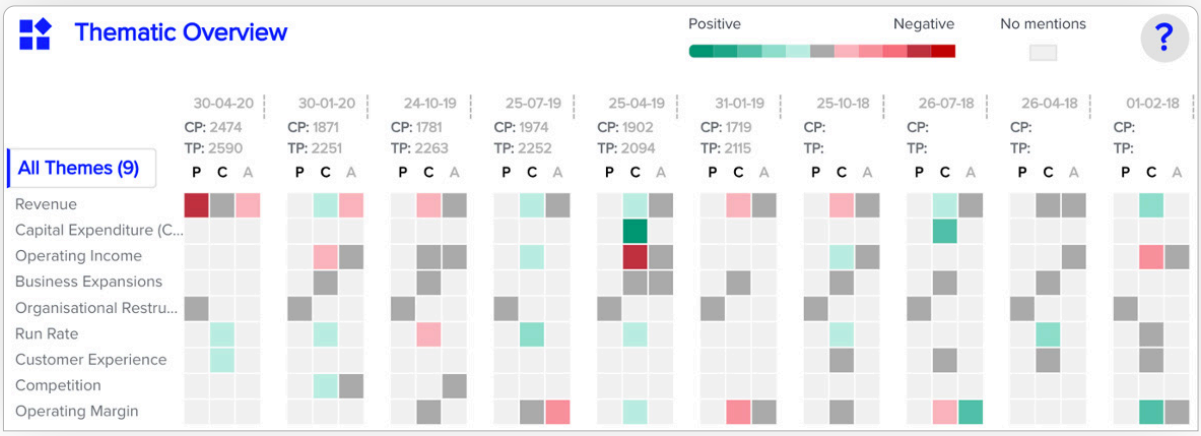


Figure 6 'Thematic Overview' pane for transcripts. Documents are split in three parts: Company presentation (P), Analysts' question in the Q&A section (A), and Company representatives' answers in the Q&A section (C).

4. SentiMine applications in finance

In the following section, we describe two typical applications of SentiMine.

Application 1: competitive comparisons

The first augments an analyst's competitive analysis capabilities by comparing a company to its seven closest peers, using the 'Peer Wizard' methodology and the 'Peer Comparison' pane (Figure 7). This methodology identifies peers by combining StarMine's Analyst Cross Coverage, Reuters Fundamentals competitor list, and industry classification using TRBC Economic Sector, Industry Sector, Industry Group and Industry.

When a user examines a company over a given time period, the 'Peer Comparison' pane identifies if the five most positive and five most negative themes in that period appear in research reports for each of the company's peers. The Peer Comparison pane also displays an aggregated sentiment score for each peer company. As shown in Figure 7, this enhanced competitive landscaping tool highlights where similar companies are doing better or worse than the company of interest.

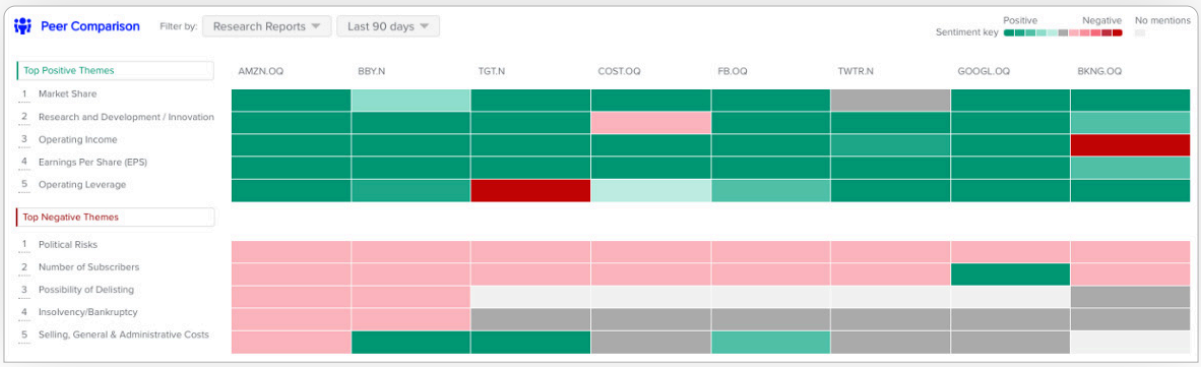


Figure 7 The 'Peer Comparison' pane allows users to compare similar companies across the same set of themes.

Application 2: changes in outlook

The second application of SentiMine pertains to the change in outlook graph (Figure 8), which highlights an analysts' sentiment on specific themes over time. A user can look at one to four themes at once and over a selected period of time to see how the sentiment changed over time, across different contributor houses or analysts.

For example, Figure 8 shows that the sentiment for 'Earnings Per Share' was slightly more positive in July and August than from September onwards and that the sentiment for the remaining themes ('Revenue', 'Regulatory Risk', and 'Price Target') has been mostly neutral for the given equity research reports.

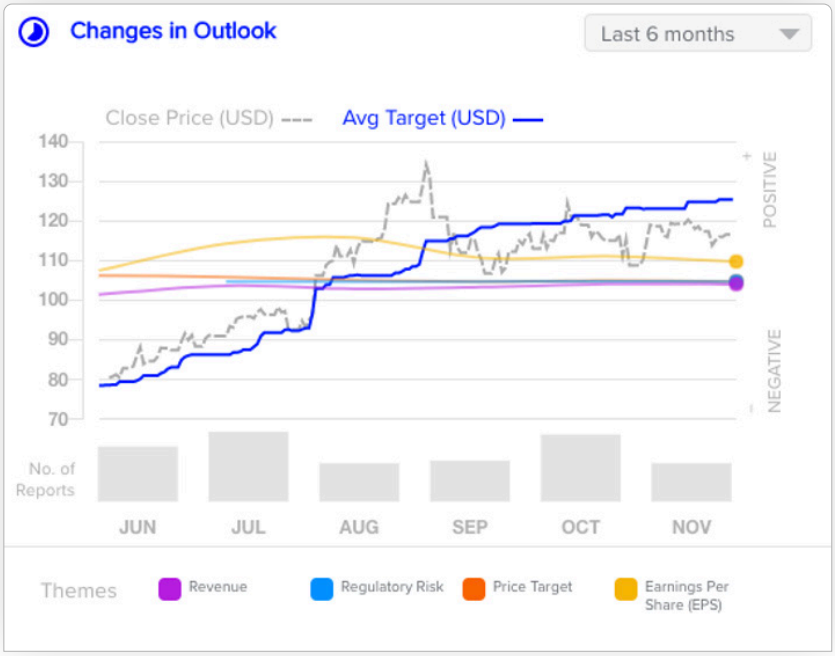


Figure 8 Changes in outlook show that the sentiment for 'Earnings Per Share' was slightly more positive in July and August than from September onwards.

5. Conclusion and future research

SentiMine is well-positioned to help analysts across finance derive more value from equity research reports, transcripts and other unstructured content by reducing the time it takes to consume content, and accelerating the identification of equity performance drivers.

Equipped with predefined sets of relevant themes, SentiMine is capable of quickly assessing theme sentiment changes over time, comparing them to the competitive landscape and observing how themes are treated during conference calls by companies and analysts.

LSEG Labs' future research on SentiMine will be focused on four core segments:

1. Improving the overall theme performance to meet at least an 80% weighted-F1 score output
2. Expanding the selection of sector-specific themes, outside of consumer retail, financial, technology and telecommunications
3. Applying SentiMine to alternative document types such as news and filings
4. Exploring alpha signals generated by SentiMine's theme sentiment

Our team strongly believe that SentiMine is rich with signals based on many factors, such as theme aggregation style, geography, sector, analysts' ratings and contributor houses. The only question is which financial firms will capitalise on the new insights powered by these emerging technologies and data science techniques.

6. Biographies

Kelvin Rocha, PhD, Lead Data Scientist, LSEG Labs Singapore

Kelvin's experience in the financial industry expands for more than 10 years and includes working as a quant researcher for a systematic stat arb hedge fund, where he was focused on building systematic equity trading models. He is also Financial Data Science Lecturer at the Yale-NUS College in Singapore.

Kai Xin Thia, MS, Senior Data Scientist, LSEG Labs Singapore

Kai Xin specialises in the research, development and deployment of large-scale machine learning systems, especially language models. He is also the co-founder of DataScience SG, a data community with 9,000+ members.

Ben Chu, MS, Senior Data Scientist, LSEG Labs Singapore

Ben has been working in the field of artificial intelligence for around 10 years, especially in the areas of NLP and knowledge representation. He is well-versed with the various frameworks for graph representation and databases, and also NLP technologies with further familiarity on linguistics and semantics.

Mihail Dungarov, MS, CFA, Senior Data Scientist, Refinitiv Workflow Intelligence

Mihail received a MSc in mathematical finance in Germany and was awarded the CFA charter in 2017. His professional career began at Deutsche Bank in Berlin and London where he has worked in capital markets, wealth management and risk roles.

Joe Whyte, Innovation Lead, LSEG Labs Singapore

Joe's experience in product stretches across computer vision, facial recognition and finance. He has spent over eight years building digital and machine learning products across a number of B2B industries.

LSEG
LABS

LSG1476474/7-21