Camille's Literature Review

1) Financial Sentiment Analysis: Techniques and Applications

https://dl.acm.org/doi/pdf/10.1145/3649451

What is the goal of the paper?

The paper is a survey whose goal is to provide a comprehensive review of financial sentiment analysis (FSA), covering both the techniques used for FSA and its applications in financial markets. It also aims to clarify the scope of FSA, define its relationship to investor sentiment and market sentiment, and propose frameworks to understand how technique and application research intersect.

Why is it important?

This paper is important because sentiment analysis in finance presents unique challenges compared to other domains, including specialized jargon, numerical and textual hybrid data, and domain-specific interpretations. Understanding and quantifying investor sentiment from sources like news, social media, and company filings can yield valuable complementary signals for forecasting stock movements, managing risk, and supporting financial decision-making. Unlike prior surveys that often focused exclusively on techniques or applications, this work connects the two, giving researchers and practitioners a more integrated understanding of FSA's role in financial tasks.

How is it solved? - Methods

Because this is a survey paper, its approach is a structured literature review rather than an experiment. The authors first define FSA's scope and distinguish between technique-driven research (datasets, models, algorithms) and application-driven research (how sentiment is used to improve financial predictions or validate economic theories). They then review major categories of FSA methods, including lexicon-based techniques, traditional machine learning, hybrid models, deep learning, and transformer-based pretrained language models, along with different levels of sentiment granularity such as sentence-level, aspect-level, and intensity scoring. The survey also reviews available datasets, feature extraction

techniques, embedding methods, and evaluation metrics. In terms of applications, the paper explores how sentiment is incorporated into predictive models for stock prices, risk, portfolio optimization, foreign exchange, and cryptocurrency, classifying data sources and highlighting trends such as the move toward implicit sentiment embeddings derived from deep models. Finally, the paper synthesizes trends, identifies gaps in the literature, and suggests open research challenges.

Results/limitations, if any.

As a survey, the "results" are synthesized insights rather than empirical outcomes. The paper highlights that FSA has evolved from lexicon-based and classical machine learning approaches to deep learning and pretrained transformer models. Applications increasingly rely on implicit sentiment embeddings rather than explicit polarity scores, as these richer representations improve predictive performance. The authors note that sentiment often provides a useful auxiliary signal for financial forecasting, with negative sentiment generally exerting a stronger and longer-lasting effect than positive sentiment. They also find that combining multiple sentiment sources, such as news and social media, can yield more robust forecasts. Limitations include poor generalizability of models across sectors and time periods due to domain drift, the high cost and scarcity of annotated financial data, the lack of interpretability in many deep models, and the difficulty of handling noisy signals and aligning sentiment with real-world market movements. The authors also note that some markets and tasks remain underexplored and that accurately quantifying nuanced sentiment, such as sarcasm or metaphor, remains a challenge.

2) Sentiment Analysis Stock Market: Sources and Challenges

https://research.aimultiple.com/sentiment-analysis-stock-market

What is the goal of the paper?

The goal of this paper is to explain what stock market sentiment analysis is, why it matters, and how it can be used to better predict stock price movements. It seeks to provide an overview of the data sources, methods, accuracy levels, and key challenges associated with applying sentiment analysis in finance, ultimately helping traders and researchers understand the practical role of sentiment in forecasting.

Why is it important?

Stock market sentiment captures the collective psychology of investors, which can significantly influence stock price fluctuations alongside traditional financial indicators. Understanding sentiment from news, social media, and financial reports allows investors to anticipate market shifts and refine trading strategies. The paper highlights research showing that incorporating sentiment data can improve price prediction accuracy by up to 20%, demonstrating its value as a complementary tool to technical and fundamental analysis.

How is it solved? - Methods

The paper describes how natural language processing (NLP) and machine learning are applied to sentiment data collected from diverse sources such as news feeds, company websites, social media platforms like Twitter and Reddit, financial reports, and economic indicators. It outlines key steps including data collection, preprocessing (tokenization, noise removal), and data labeling into positive, negative, or neutral categories to train machine learning models. Methods include rule-based systems, lexicon-based techniques, and advanced machine learning models like XGBoost and BERT. BERT, in particular, is highlighted as a powerful model that uses context-aware embeddings and attention mechanisms to extract investor sentiment from text, achieving over 97% accuracy in experiments.

Results/limitations, if any.

The article cites multiple studies showing that sentiment analysis can predict stock price movements with accuracy ranging from 60% to 99%, depending on data size, model sophistication, and market context. Combining sentiment with technical and fundamental indicators improves prediction reliability, and large language models like GPT have demonstrated strong portfolio returns with high Sharpe ratios. However, challenges remain, including filtering noisy data, handling the evolving nature of language, integrating qualitative sentiment with quantitative financial metrics, and maintaining compliance with privacy regulations. The authors emphasize that sentiment analysis should be viewed as a complementary tool rather than a standalone predictor and that models require continual fine-tuning to remain effective as market conditions and investor behavior shift.

3) Stock Market Forecasting Based on Text Mining Technology: A Support Vector Machine Method

https://arxiv.org/abs/1909.12789

What is the goal of the paper?

The goal of this study is to investigate whether text mining combined with support vector machine (SVM) models can improve the prediction of Chinese stock market trends and prices. The authors aim to show that integrating online news data with market data produces more accurate forecasting models, helping reveal how investor sentiment affects short-term price movements.

Why is it important?

Predicting stock markets is challenging due to their volatility and dependence on numerous economic and psychological factors. In China, where most investors are retail traders with limited financial knowledge, external information such as news can heavily influence decisions. This paper is significant because it demonstrates how text mining can be systematically applied to large-scale news datasets to enhance price prediction, providing valuable insight for regulators, traders, and researchers studying emerging markets.

How is it solved? - Methods

The study collected over 2.3 million Chinese online financial news articles from 2008 to 2015 and combined them with stock price data for 20 highly traded Chinese stocks. The authors developed domain-specific stop-word and sentiment dictionaries, scoring words on a scale from –5 to +5 to quantify sentiment. Using these dictionaries, they constructed daily input vectors combining text sentiment features with market variables (adjusted close price, time-lagged data). They trained both support vector regression (SVR) models for price prediction and support vector classification (SVC) models for trend prediction, using LIBSVM with polynomial and sigmoid kernels. Parameter optimization was performed using grid search and genetic algorithm-based techniques to find optimal C and γ values, with additional experiments testing time lag effects and input data expansion strategies.

Results/limitations, if any.

The SVR model achieved excellent performance, with a strong correlation coefficient (SCC ≈ 98.5%) and low mean squared error (MSE ≈ 0.00328) for price predictions. It accurately captured sharp price fluctuations, although with a one- to two-day delay. SVC models achieved about 59% classification accuracy, consistent with results in similar research, but showing room for improvement. The study also found that news impacts stock prices for less than two days and that parameter γ plays the most critical role in model performance. Additionally, the authors proposed a novel approach to calculate news source impact factors, revealing that traditional media sources and large-audience platforms exert the greatest influence on investors. Limitations include weaker results when news volume was low, partial sentiment dictionaries that may bias inputs, and the need for more extensive text sources and standardized sentiment scoring to improve generalization.

4) Using News to Predict Investor Sentiment: Based on SVM Model

https://www.sciencedirect.com/science/article/pii/S187705092031588X?via%3Dihub

What is the goal of the paper?

The goal of this paper is to explore how news articles influence investor sentiment and to build a model that predicts changes in investor sentiment using a support vector machine (SVM). The authors focus on the psychology line (PSY) as a sentiment indicator and aim to demonstrate that SVM-based models can accurately forecast investor mood and thereby improve stock market decision-making.

Why is it important?

This work is important because most prior studies have concentrated on how news affects stock prices directly, without considering the intermediate role of investor emotions in market movements. By focusing on investor sentiment, the paper highlights a crucial factor that drives price fluctuations and trading behavior. The research emphasizes that better sentiment prediction could guide investors toward more rational decision-making and potentially lead to higher cumulative returns in the market.

How is it solved? - Methods

The researchers collected over 19,000 company news texts and 10,000 industry news texts from December 2013 to March 2017, focusing on the medical industry to control for sector-specific effects. After cleaning and segmenting the text using Python's jieba library, they vectorized the data using a doc2vector model. PSY indicators and corresponding stock open/close prices were gathered from the Wind database. The processed data were labeled based on whether PSY was above or below 50 and fed into the SVM model, which was trained to classify sentiment as positive or negative. The authors also tested a universum SVM (U-SVM) model to incorporate industry news and designed comparative experiments to analyze the effect of different data sizes and alternative indicators (e.g., price ratio vs. PSY).

Results/limitations, if any.

The SVM model achieved a prediction accuracy of about 59% when using PSY as the target indicator, outperforming other configurations. Results showed that adding industry-level news slightly reduced accuracy due to overfitting, suggesting that firm-specific news is a stronger predictor of investor sentiment. Accuracy also declined slightly when data volume was reduced, reinforcing the importance of using comprehensive datasets. While the model proved effective in capturing sentiment trends, limitations include its relatively modest accuracy (just under 60%), the narrow sectoral focus (medical industry only), and potential overfitting when incorporating too much external data. Future work could expand to multi-sector data and explore hybrid models to improve robustness.

5) Twitter Sentiment Geographical Index Dataset

https://www.nature.com/articles/s41597-023-02572-7

What is the goal of the paper?

The goal of this work is to introduce the Twitter Sentiment Geographical Index (tSGI), the most comprehensive open-source dataset of location-specific sentiment derived from over 4.3 billion geotagged tweets since 2019. The authors aim to provide researchers with a

high-resolution, multilingual, global database that captures expressed sentiment at the daily and county/city level, allowing for analysis of social well-being trends, responses to major events, and cross-country comparisons.

Why is it important?

Traditional measures of subjective well-being (SWB), such as surveys, are limited by high costs, low frequency, and delayed results. The tSGI dataset addresses these challenges by offering near real-time, large-scale sentiment indicators, which are crucial for policymakers and researchers monitoring well-being trends, public reactions to crises, and policy impacts. By leveraging social media data, the dataset allows for finer spatial and temporal granularity, helping to complement conventional economic and health metrics with real-time behavioral signals.

How is it solved? - Methods

The dataset was constructed using the Harvard CGA Geotweet Archive, collecting tweets with geolocation attributes from 2019 onward. Text data underwent extensive preprocessing, including URL and emoji removal, user mention normalization, and truncation to a maximum of 52 words. Sentiment was computed using a multilingual Sentence-BERT (S-BERT) model, which generated contextual embeddings for over 50 languages. These embeddings were fed into a four-layer neural network classifier trained on the Sentiment140 dataset (1.6M labeled tweets), achieving 83% test accuracy. Each tweet was assigned a sentiment probability score using a SoftMax layer, and aggregated sentiment scores were calculated at multiple spatial levels (global, country, state, and county/city) on a daily basis.

Results/limitations, if any.

The tSGI dataset offers coverage across 164 countries, with an average of nearly 3 million geotagged tweets processed per day. The model significantly outperforms dictionary-based approaches like LIWC and VADER, achieving higher accuracy and F1 scores (0.829). Validation shows a strong correlation between tSGI sentiment indices and independent well-being measures, such as the Hedonometer Happiness Index, confirming the dataset's reliability. However, the reliance on geotagged tweets—only 1–2% of all tweets—may introduce sampling bias, as geotagged posts tend to be more positive. Additionally, the dataset represents Twitter users rather than the entire population, and socio-demographic

information is unavailable, limiting representativeness. Despite these challenges, tSGI remains a valuable and timely proxy for monitoring global sentiment and well-being.

6) A Comparative Study of Machine Learning Algorithms for Stock Price Prediction Using Insider Trading Data

https://arxiv.org/abs/2502.08728v2

What is the goal of the paper?

The goal of this research is to empirically evaluate multiple machine learning algorithms—Decision Trees, Random Forests, Support Vector Machines (SVM) with different kernels, and K-Means Clustering—for predicting stock prices based on insider trading data. The study aims to identify which algorithms deliver the most accurate results while balancing computational efficiency and to determine which features of insider trading activity are most predictive of future stock price movements.

Why is it important?

Insider trading activity offers a unique window into company sentiment and expectations, potentially signaling price movements before they appear in the market. Understanding which machine learning algorithms perform best on insider trading data can guide investors and financial analysts toward better-informed decisions. This work is significant because it highlights the potential of combining alternative data sources with predictive analytics, contributing to the growing field of data-driven finance and algorithmic trading.

How is it solved? - Methods

The study collected Tesla insider trading data from April 2020 to March 2023 using the Finnhub API, resulting in 1,997 samples of transactions. After cleaning and preprocessing the data, the researchers engineered two new features—Dollar Volume and Transaction Type—to enhance predictive power. Recursive Feature Elimination (RFE) was applied to select the four most relevant features: Shares, Transaction Date, Dollar Volume, and Type. The dataset was split into training (70%) and testing (30%) sets. Each algorithm (Decision Tree, Random Forest, SVM with linear/polynomial/RBF kernels, and K-Means Clustering)

was trained and evaluated using accuracy as the primary metric and runtime as a secondary metric to capture computational efficiency.

Results/limitations, if any.

SVM with the RBF kernel achieved the highest prediction accuracy at 88%, followed by Random Forest at 83%, though both models had longer runtimes (28 and 18 minutes, respectively). Decision Trees were the fastest but least accurate at 68%, while K-Means Clustering achieved 73% accuracy. The study concluded that higher computational cost tends to correlate with better accuracy, making SVM-RBF the most suitable choice for this dataset despite its longer runtime. Limitations include the relatively small dataset (limited to Tesla transactions) and reliance solely on insider trading data, which may omit other influential market drivers such as macroeconomic conditions, earnings reports, or news sentiment. The authors suggest integrating additional data sources to further improve predictive performance and generalization.