Camille Esteves
IDC6940
9/22/2025

Summaries

Research Paper 1: Twitter Sentiment Geographical Index Dataset
Authors: Yuchen Chai, Devika Kakkar, Juan Palacios, Siqi Zheng
Year: 2023

The Twitter Sentiment Geographical Index is a massive dataset built from roughly 4.3 billion geotagged tweets posted worldwide since 2019. It captures expressed sentiment — positive, negative, or neutral — at a daily frequency and at fine geographical granularity, going as detailed as counties or cities across 164 countries. This makes it a powerful dataset for anyone trying to model how sentiment shifts across both time and location. For my project, it could be really valuable because I can train an SVM to generalize sentiment across regions or even compare patterns between locations. The daily frequency is also perfect for aligning with other time series data, like stock prices, to analyze possible correlations. That said, tweets are notoriously messy, so I'd need to handle text preprocessing, translation for multilingual tweets, and deal with imbalances where neutral sentiment often dominates. Plus, geolocation data is missing for many tweets, so coverage won't always be complete — something to keep in mind when aligning with other datasets.

Research Paper 2: A Comparative Study of Machine Learning Algorithms for Stock Price Prediction Using Insider Trading Data
Authors: Amitabh Chakravorty, Nelly Elsayed
Year: 2025

This paper explores how insider trading data — when company executives and insiders buy or sell their own company's stock — can be used to predict price movements. The study focuses on Tesla stock between April 2020 and March 2023 and compares multiple algorithms, including decision trees, random forests, and several SVM kernels. Interestingly, they find that SVM with an RBF kernel performs the best among the tested models, which gives me a great benchmark for my own work. The paper also uses feature selection techniques like Recursive Feature Elimination to reduce dimensionality and avoid overfitting, which could be helpful to replicate in my project. However, insider trading data tends to be sparse and irregular, so aligning it with daily price data is tricky. There's also a risk of overfitting since insider trades sometimes reflect information that isn't public and may not generalize well to other stocks or time periods.