Отчет по статье Stock Prediction Using Twitter Sentiment Analysis(Anshul Mittal, Arpit Goel)

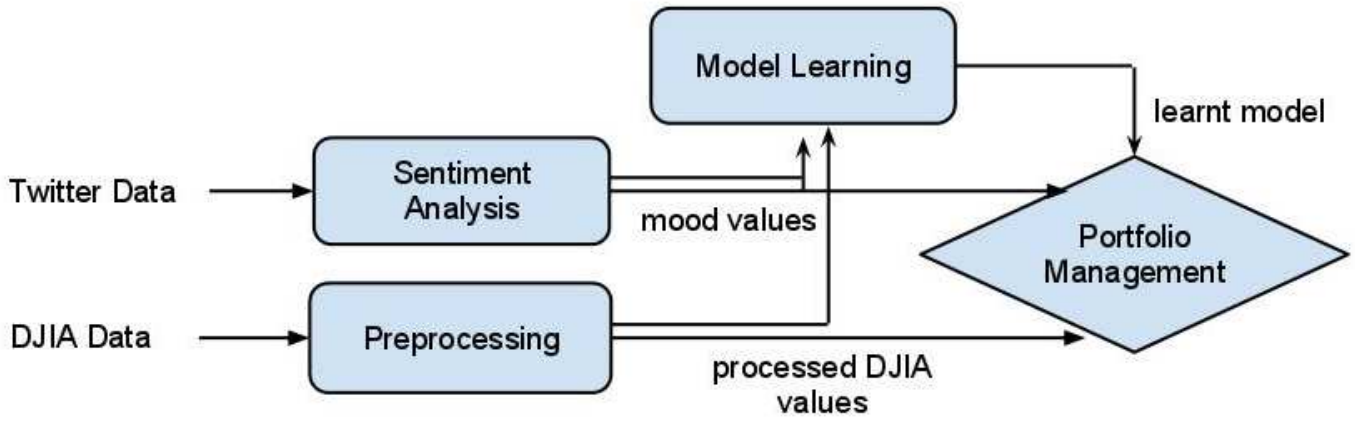
Идея:

Предсказание трендов движения рынка DJIA, используя данные из Твиттера и предыдущие данные рынка. Данные твитов используются для того, чтобы предсказать публичное настроение людей(все твиты распределяются по 4 классам: Calm, Happy, Alert и Kind), и далее полученный тип настроения и предыдущие данные рынка подаются на вход SOFNN(Self Organizing Fuzzy Neural Network) для предсказания будущего поведения. На основе модели авторы также построили упрощенную стратегию построения портфолио, принимающую решения о покупке и продаже.

Данные:

Авторы использовали ежедневные данные DJIA(open, close, high, low) в период с 06.2009 по 12.2009 и около 476 миллионов публично доступных твитов(с указанием timestamp, имени пользователя и текста твита) на тот же период. В папке данной статьи в нашем репозитории написано, что такие данные твитов более публично недоступны, однако есть доступные данные по 1600000 твитам в период с 6.04.2009 по 16.06.2009 с сайта http://help.sentiment140.com/for-students. Они просят цитировать их при использовании этих данных, но думаю это не проблема.

Использование данных:



Препроцессинг: пропущенные данные DJIA (выходные и праздничные дни) восполняются вогнутой функцией(если x и y - два имеющихся значения, то первому пропущенному дню между ними присваивается значение (y+x)/2, далее ((y+x)/2 + y)/2 и так далее до заполнения всех пропусков). Такой подход, говорят авторы, оправдан, так как данные рынка обычно следуют (за исключением аномальных возрастаний и падений) вогнутой функции. Наиболее крутые падения и возрастания сглаживаются(без нарушения направления тренда). Периоды наиболее изменчивой активности(volatile activity) удалены как из тренировочной, так и из тестовой выборки, ввиду того что их очень тяжело предсказывать. Все данные приведены к одному рангу с помощью нормализации z-score.

Работа с твитами:
Общая схема: генерируется лист из 65 слов, отвечающих 6 настроениям(согласно Profile of Mood States (POMS) questionnaire), этот список расширяется с помощью синонимов. Все твиты фильтруются и оставляются только те, которые скорее всего выражают чувства(содержат фразы типа I'm feeling, I am и так далее). Далее для каждого слова из POMS для каждого дня считается
score = (# раз, которое слово(или синоним) встречалось в твитах этого дня)/(количество попаданий для всех слов).
Эти результаты классифицируются по 6 настроениями согласно инструкциям POMS, и затем 6 настроений из POMS перераспределяются по 4 классам настроений, выбранных авторами.
Для того, чтобы понять, какие из 4 настроений и насколько влияют на предсказание рынка применяется причинность по Грэнджеру(в результате выяснилось, что лучшая комбинация это Calm + Happy).

Обучение и предсказание:
Для обучения и предсказания были опробованы модели линейной регрессии, логистической регрессии, SVM и SOFNN. Делались предсказания как просто направления движения рынка, так и точного его значения. Для проверки того, как обученная модель обобщается на тестовые данные, авторы придумали делать последовательную кросс-валидацию: тренировка идет по всем дням до какого-то определенного дня, а тестирование на k днях строго после этого определенного дня. Этот подход более разумен в этом случае, чем обычная кросс-валидация, потому что здесь нам не смысла предсказывать прошлые результаты по данным будущего.

Результаты:
Наилучший результат был показан при обучении SOFNN на признаках Calm, Happy и предыдущие значения DJIA - 75,56% accuracy. Построенная на основе предсказаний этой модели стратегия покупки и продажи, как показывают авторы, вроде бы даже зарабатывает определенное количество Dow points.

Дополнительно:
1) Статья во многом опирается на статью J. Bollen and H. Mao. Twitter mood as a stock market predictor(http://arxiv.org/pdf/1010.3003.pdf), используя похожие алгоритмы обучения, но другие методы обработки твитов и тестирования результатов. Для полноты картины

наверное стоит изучить и эту статью.

Воспроизведение и вывод:
Для того, чтобы воспроизвести результаты, нужно разобраться с тем, что такое SOFNN и как она работает. В списке ссылок приведена ссылку на статью G. P. Gang Leng and T. M. McGinnity. An on-line algorithm for creating self-organizing fuzzy neural networks (http://www.sciencedirect.com/science/article/pii/S0893608004001698). Также есть соображение, что возможно в нашей стране твиттер используется не так активно, и для реализации возможно нужно использовать данные более распространенных социальных сетей, или, может быть, анализировать посты финансовых блогов и новостей.