Stock Prediction by Polyglot Sentiment Analysis on Twitter

Yizhou Wang

Marianopolis College, Canada

ABSTRACT

Research in the economics field has extensively documented the impact of media sentiments on the stock market. Sentiment analysis, as a tool to predict equity prices, has been popularized in the past years. Recently, Twitter has received a lot of attention due to the diversity of opinions on social media platforms. A common obstruction to sentiment analysis is the resource gap between English and other languages. This pilot study examines the effect of polyglotism in tweets concerning bilingual companies and develops a model that extracts sentiments from polyglot tweets to make price predictions. Results suggest that taking non-English tweets into consideration decreases errors in price predictions and that the random forest models have higher performance than linear regression models. The results of this pilot study need to be confirmed with larger sets of data.

Introduction

The problem of stock market prediction is one that is extensively studied in the economical field. The emergence of behavioral economics offers a new perspective to modern economics because individuals are no longer modeled as entirely rational beings (what is called the homo economicus). In reality, people's decision-making ability is limited by the complexity of the decision, the cognitive limitations of the human mind, and the amount of time available. Sentiments affect investors' decisions when their expectations for the market deviate from what objective market information allows to infer. In particular, media pessimism/optimism is observed when media (newspapers, social platforms) express opinions that are unfavorable/favorable towards a certain company or stock. Extensive research has studied the relationship between media sentiment and market movements.

Media Sentiments and Market Trends

Paul C. Tetlock (2005) is among the first to search for evidence that media content can predict movement in broad market indicators such as trading volume and [...]. Tetlock analyzed content from a Wall Street Journal column that includes market reports, post-mortems, and predictions by professional analysts. Tetlock counted the words in 77 General Inquirer categories and performed a principal component factor analysis to obtain a single factor capturing the maximum variance in the categories. This factor turns out to be strongly related to the level of media pessimism. Tetlock concluded that 1) following high levels of media pessimism, prices move downwards and eventually revert to the fundamentals, 2) following unusually high or low levels of media pessimism, high trading volume is observed, and 3) downward prices forecast high levels of media pessimism. (Tetlock, 2005)

Mendoza et al. (2022) suggested that negative sentiments affect prices more intensely than positive sentiments (Mendoza-Urdiales et al., 2022). This phenomenon can be explained by a behavioral economics concept called loss aversion, a cognitive bias that states that the pain of losing is psychologically more intense than the pleasure of gaining an equivalent amount.

Sentiment Analysis

Natural Language Processing (NLP) is a field of Artificial Intelligence that involves extracting, from natural human language, structured information intelligible to machines. Sentiment analysis is an NLP technique used to determine a text's subjectivity and attitude. Sentiment analysis can be used to 1) determine a text's polarity, that is, whether it is positive, neutral, or negative, and 2) detect emotions such as happiness, sorrow, and nervousness. Multilingual sentiment analysis requires a considerable amount of resources, such as sentiment lexicons, translations, and/or noise detection algorithms.

Semantic Differential

Semantics differential (SD) is a measurement scale for the subjective perception of a concept on a set of bipolar scales. It is often visualized as a set of interval scales, each of which situates the concept between two opposite adjectives, such as "bright-dark", "sweet-bitter", and "beautiful-ugly". Among all scaling techniques, SD is considered relatively reliable for assessing respondents' opinions and therefore can be used as a proxy for media attitude.

Sentiment Analysis on Twitter

In the past few years, much attention has been accorded to the social platform Twitter. Unlike columns written by professionals, content on Twitter represents the attitudes of a larger public. The platform is promising in terms of polarity extraction as people around the globe express their opinions on a diversity of topics, including politics, environmental issues, emerging technologies, and economic events.

Due to the form of textual data on Twitter, sentiment analysis presents a unique set of challenges. As noted by Tellez et al., tweets are short and informal; they contain slang, grammatical errors, and typos (Tellez et al., 2016). Misspelled words and invalid syntax may result in detection failure, in which case sentiments would be lost. Therefore, auto-correcting software and noise-removing algorithms are likely required.

Another consequence of Twitter's informal register is intratextual polyglotism. A single tweet can incorporate words from two or more languages. Currently, a wide gap exists between English models and other languages (Tellez et al.). This paper analyzes comments about bilingual companies; hence, the text data are expected to contain English and French. Differences between these two languages are to be considered. English is a Germanic language, whereas French is a Romance language. Here, the Google Translate API is used to convert all tweets into English before conducting sentiment analysis. Another feature that distinguishes French from English is the presence of diacritic symbols (accents, tremas, cedillas). In informal environments, their incorrect usage is one of the most important sources of spelling errors (Tellez et al.). Here, diacritic symbols are removed using the Unicode library in Python.

Research Question

In this pilot study, sentiment analysis is performed on bilingual companies, here defined as companies that headquarter themselves in bilingual regions, operate in two or more languages, and serve customers that speak two or more languages. The goal is to answer the following questions concerning equity prices of bilingual companies:

1. What portion of tweets regarding these companies are non-English?
2. How does taking polyglotism into consideration affect prediction accuracy?
3. Which machine learning model (Linear Regression or Random Forest) is more performant at predicting price change from polarity score?

Experimental Design

This paper proposes a pilot study that considers tweets addressed to a bilingual company over a range of 35 days. The company in study is Desjardins Group. Desjardins is a financial service cooperative whose products include chequing accounts, insurance, and stock brokerage. The company's headquarter is in Montreal, where 53.9% of the population speaks both English and French, 37.0% speaks French only, and 7.4% speaks English only (Statistics Canada, 2011). First, the prevalence of polyglotism is examined. A search with query "@" or "#" + "*The company name"* is made without discriminating the language. In the search results, the volume of tweets in each language (French, English, and others) is analyzed.

Then, the present paper examines how public sentiments on Twitter affect the stock market. Polyglot tweets data are first converted to uniform English text data using the Google Translate API. The polarity of the tweets is extracted using the Vader Lexicon, a sentiment analysis lexicon specially trained to capture sentiments expressed on social media. The last part of the study consists in training a model that predicts price movement based on polarity scores. Mean square errors are analyzed to compare the performance of different models.

Methods

Collection of Data

Tweets are collected via the Twitter API. A Twitter developer account is registered, and an app is set up. The free API version only gives access to tweets from at most seven days ago and is limited to 100 tweets maximum per search. These limitations have to be accounted for in order to gather a statistically significant amount of data. Tweets are gathered periodically (once per week) from late November 2022 to late December 2022. Since 100 tweets do not make data for the full week, a Python program is written to repeatedly search tweets dating up to the oldest tweet found in the previous search. The program also uses the Regular Expression module in Python to remove noise (punctuation marks and diacritic symbols).

Historical data of three Québec-based companies (Desjardins, Bell, and LightspeedHQ) are downloaded from Yahoo Finance. The data are CSV files that contain the nominal closing price and the adjusted closing price for each trading day. In the adjusted closing price, dividends are subtracted (as the true value of a stock decreases after paying the dividends), and the price is readjusted when the company splits its stocks (as the nominal price is divided by the split ratio). The adjusted closing price is used because it compensates for artificial price changes and thus depicts price movements more accurately.

Polarity Mining

The Vader Lexicon is used to conduct sentiment analysis on the collected tweets. Vader stands for Valence Aware Dictionary and Sentiment Reasoner and is specifically optimized for social media. The goal is to obtain a polarity score between [-1, 1] is extracted for each trading day.

When a tweet is given as input, Vader returns the weight of each sentiment (nonnegative values that sum to 1): positive, negative, and neutral. A compound score, between [-1, 1], is also returned. Here, the pair consisting of the positive weight and the negative weight is used as the independent variable.

Correlation and Prediction

Models are trained to predict the direction of the stock market based on tweets published between two trading day. More specifically, the models predict the dependant variable, the price change, given the independent variables, the

weight of positive emotion and the weight of negative emotion. Here, two different techniques are tested: Linear Regression and Random Forest Regression. The different models that stem from them are compared in performance. Performance is quantified by the mean square error and the number of days where the model correctly predicted the direction of the change. The mean square error formula is shown below.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

In the control experiment, tweets are entered into the sentiment analyzer without translation into homogeneous English. Therefore, tweets in French and other languages are not considered. In the main experiment, tweets are first translated into English using the Googletrans API. Therefore, all tweets are considered by the sentiment analyzer disregarding their original languages. An improved performance exhibited by the polyglot model compared to the English-only model would imply that non-English tweets have a significant impact on stock price changes of bilingual companies. This would indicate direction for future research, as state-of-the-art models can be improved by better capturing sentiments contained in non-English texts.

In summary, four models are obtained for each company: Linear Regression-untranslated, Linear Regression-translated, Random Forest-untranslated, and Random Forest-translated. Performance is compared in order to determine the best machine learning technique and the impact of polyglotism.

Results

Tweets

Tweets addressing three companies are collected. Tweets range from November 21 to December 23, 2022. All tweets between two consecutive trading days are assembled and are assumed to be responsible for the price change between these two days. Price change on day X is calculated as the difference between the close price on day X and the close price on day X-1. Volume of tweets, measured in number of words, is also recorded.

Below is an example for Deajardins, in the month of November.

Table 1: Time Series of Price Change and Tweets for Desjardins from 11/18 to 11/30

	Date	Close	Change	Tweets	Volume
0	11/18/2022	9.891	0		0
1	11/21/2022	9.915	0.024	Learn how our critical illness insurance can make hard times easier and protect you and your loved ones Click here to request a free insurance quote Desjardins Toronto GTA httpstcooFFQArA77I httpstcolzA2l9pGmu Releve de compte Desjardins et autres documents en ligne Suivre mes operations bancaires et mon contrat relevedecompte desjardins application httpstcosKvdlcNfLp RT ThomasRolain La fuite des donnees des clients de la banque Desjardins est arrivee en 2019 Depuis lors toujours aucune mesure de pro	75
2	11/22/2022	9.963	0.048	Tres fier de voir beaucoup de noms de nos collegues de VMD Desjardins Bravo httpstcoiCDOHjfpO7	15
3	11/23/2022	10.004	0.041	Arnaud Desjardins e Mason Nyhus dois quarterbacks em uma classe propria na Vanier Cup httpstcopX6RZbLOUs Esportes Arnaud classe Cup Desjardins Take a	43

	Date	Close	Change	Tweets	Volume
				look Proud to be associated with this amazing organization The only topranked Canadian company femalefriendly organization proud canadianbusiness desjardins topranked httpstco5BjxJoRC49	
4	11/24/2022	10.038	0.034	RT WomensArtForAll Claire Desjardins WomensArt ArtByWomen WomenArtists BreakfastFor details see the listing on Claires Insta Claire Desjardins WomensArt ArtByWomen WomenArtists BreakfastFor details see the listing on Claires Insta httpstco33oWz1EbgI or visit httpstcoXqqvkoX1Er httpstcoBigmcDx0U3 RT TorontoStar Armine Yalnizyan discusses current economic painsdesjardins responsibleinvesting partner financialliteracy investing Merci a la Caisse Desjardins de MercierEstAnjou qui a genereusement offert des cadeaux pour le temps des Fetes a nos employes afin de souligner leur devouement aupres des personnes handicapees Desjardins Employe httpstcoBp5uEH4dGi Armine Yalnizyan and Darryl Brown provide guidance on navigating inflation and the looming recessiondesjardins responsibleinvesting partner financialliteracy investing Armine Yalnizyan discusses current economic painsdesjardins responsibleinvesting partner financialliteracy investing Armine Yalnizyan and Darryl Brown provide guidance on navigating inflation and the looming recessiondesjardins responsibleinvesting partner financialliteracy investing Armine Yalnizyan discusses current economic painsdesjardins responsibleinvesting partner financialliteracy investing RT funimag Savezvous que la tour du Stade Olympique de Montreal possede un 2eme ascenseur incline parallele a celui qui est exterieur I Savezvous que la tour du Stade Olympique de Montreal possede un 2eme ascenseur incline parallele a celui qui est exterieur Il sert de monte charge avec une pente allant jusqua 196 funiculaire StadeOlympique Montreal Quebec Desjardins parcolympique Montreal httpstco8PHU9amgpo	201
5	11/25/2022	10.052	0.014	RT TorontoStar Armine Yalnizyan discusses current economic painsdesjardins responsibleinvesting partner financialliteracy investing RT TorontoStar Armine Yalnizyan discusses current economic painsdesjardins responsibleinvesting partner financialliteracy investing RT WomensArtForAll Claire Desjardins WomensArt ArtByWomen WomenArtists BreakfastFor details see the listing on Claires Insta	57
6	11/28/2022	10.036	-0.016	RT TorontoStar Armine Yalnizyan and Darryl Brown provide guidance on navigating inflation and the looming recessiondesjardins respons Gilles Farcet oeuvre pour la transmission spirituelle au sein de la lignee dArnaud DesjardinshttpstcoSULOIYh9xunondualite Prajnanpad Desjardins Farcet httpstcoSU8TL1BLyu	35
7	11/29/2022	10.014	-0.022	Balenciaga suing production company for 25 million over controversial campaign desjardins balenciaga nicholasdes des jardins jardinsllc httpstcoHc86dzccpV Quelle est la difference entre echanger et racheter ses actions CRCD httpstcoAaUuIwIHcAfinancespersonnelles retraite101 retraite independancefinanciere investir investissement investisseur investing CRCD CapitalRegional Desjardins	102
8	11/30/2022	10.096	0.082	Il canadese volanteTutte le parate piu belle di Desjardins contro il Sangiuliano CityForzaNovara NovaraFC How do you modernize an oldbuilding from 1878 and turn it into a modern financial institution with awardwinning interiordesign and stateoftheart specializedfurniture Simple Just copy what Desjardins did with Le Windsor httpstcoSQCdMLNpK0 httpstcoVgxCaMTuEn RT Grenieremplois Montreal ou LevisDesjardins cherche a combler un poste de gestionnaire de campagnes multiplateformeshttpstco RT TorontoStar Armine Yalnizyan and Darryl Brown provide guidance on navigating inflation and the looming recessiondesjardins respons RT TorontoStar Armine Yalnizyan and Darryl Brown provide guidance on navigating inflation and the looming recessiondesjardins respons	134

In the 8-day sample above, non-English sentences are underlined (13 in total). Most are French (11 sentences), with Spanish appearing occasionally (2 sentences). Misspellings are also frequent. Misspelled words are skipped over by the sentiment analyzer; hence misspelling can cause polarity to shift towards neutral.

Polarity Mining

Polarity is extracted from each day's tweets. Two arrays of polarity scores are obtained, one from the untranslated dataset and another from the translated dataset. Sample data for the first 8 trading days are shown below.

Table 2. Polarity of Untranslated Tweets

	Date	Comp	Negative	Neutral	Positive
0	11/18/2022	0	0	0	0
1	11/21/2022	0.802	0.075	0.776	0.148
2	11/22/2022	0	0	1	0
3	11/23/2022	0.893	0	0.78	0.22
4	11/24/2022	-0.25	0.016	0.984	0
5	11/25/2022	-0.128	0.027	0.973	0
6	11/28/2022	-0.128	0.042	0.958	0
7	11/29/2022	-0.599	0.067	0.933	0
8	11/30/2022	-0.459	0.045	0.955	0

Table 3: Polarity of Translated Tweets

	Date	Comp	Negative	Neutral	Positive
0	11/18/2022	0	0	0	0
1	11/21/2022	0.5574	0.138	0.704	0.158
2	11/22/2022	0.5256	0	0.78	0.22

	Date	Comp	Negative	Neutral	Positive
3	11/23/2022	0.8934	0	0.78	0.22
4	11/24/2022	-0.25	0.016	0.984	0
5	11/25/2022	-0.128	0.027	0.973	0
6	11/28/2022	-0.128	0.042	0.958	0
7	11/29/2022	0	0	1	0
8	11/30/2022	-0.456	0.045	0.955	0

For each dataset, the weights of positive, negative, and neutral emotions are illustrated in a pie chart. In both case, neutral emotion prevails (90% in the untranslated set and 87% in the translated set), though the translated set captured an extra 2% of polarized (either positive or negative) emotions.

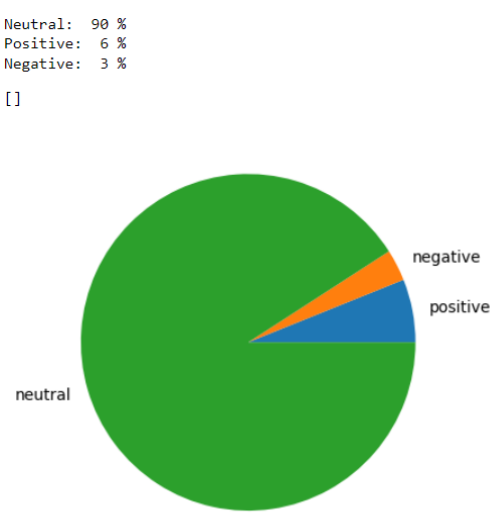

Figure 1. Polarity of Untranslated Tweets

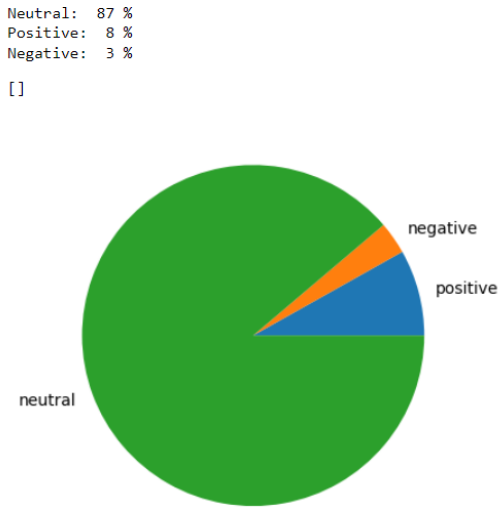

Figure 2. Polarity of Translated Tweets

Prediction

First, two prediction models are trained using Linear Regression. Graphs of predicted and actual price changes are shown below.

Figure 3. Prediction using Linear Regression and untranslated tweets

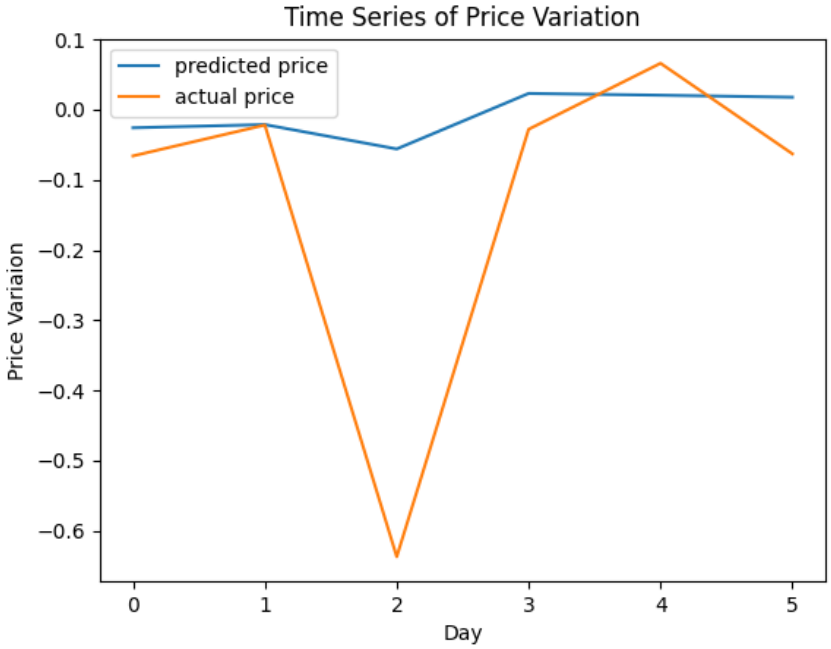

Figure 4. Prediction using Linear Regression and translated tweets

Mean square error is calculated for both models. The untranslated model has a mean square error of 0.0787, whereas the translated model has a mean square error of 0.0584. In the period of 6 trading days, the untranslated model correctly predicted the direction (sign) of the change twice, whereas the translated model correctly predicted the direction four times.

Then, two different prediction models are trained using Random Forest. Graphs of predicted and actual price changes are shown below.

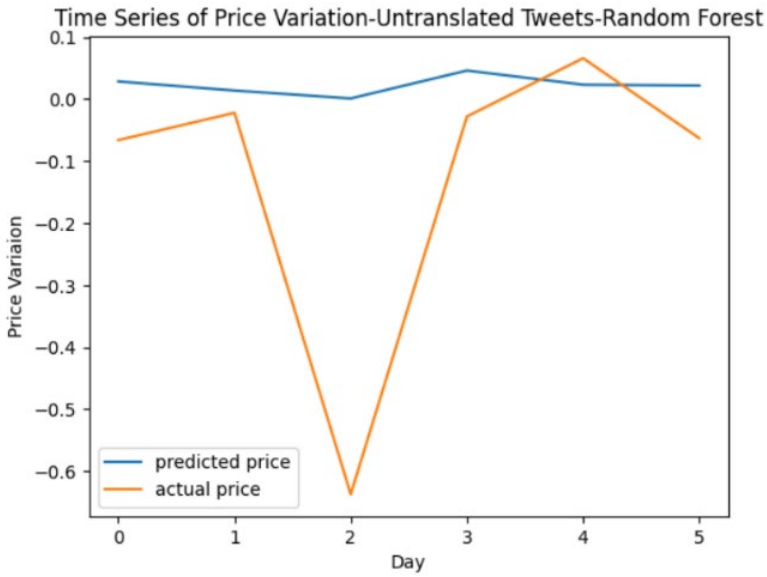

Figure 5. Prediction using Random Forest and untranslated tweets

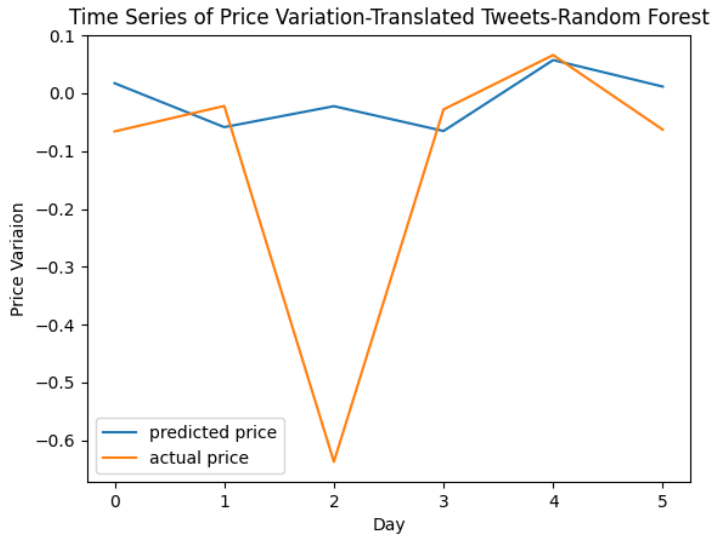

Figure 6. Prediction using Random Forest and untranslated tweets

Mean square error is calculated for both models. The untranslated model has a mean square error of 0.0642, whereas the translated model has a mean square error of 0.0555. In the period of 6 trading days, the untranslated model correctly predicted the direction (sign) of the change twice, whereas the translated model correctly predicted the direction four times.

Discussion

The results suggest that taking polyglotism into consideration improves prediction accuracy. Mean square error is reduced by 25.8% for the Linear Regression model, and by 13.6% for the Random Forest model.

All predictions are within 0.10$ from the actual change, except for day 2 where the models failed to anticipate the striking drop of -0.60$. The mean square errors (between 0.0555 and 0.0787) are significant; hence, the models are inaccurate when predicting the exact price change.

Nonetheless, in practical contexts, correctly predicting the direction of change (whether the price increases or decreases) can already provide useful information about the market. For both machine learning models, the translated model correctly predicted the direction of change in four of the six days, whereas the untranslated model correctly predicted the direction in two of the six days. This again suggests that homogenizing into English prior to training improves accuracy.

The results also suggest that Random Forest is superior in performance to Linear Regression in the task of predicting price change from sentiment scores. For untranslated tweets, the mean square error was 18.4% lower in the Random Forest model; for translated tweets, the mean square error was 5.0% lower.

Limitations

Due to Twitter API's time limit, only tweets less than one week old are available. Despite collecting data weekly for five weeks, only 35 days of tweets are available, which amount to around 25 trading days for the training and the test sets combined. Hence, conclusions made in the discussion section need to be verified with larger datasets. In subsequent research, the elevated version of the API can be used to collect tweets over a longer period.

Other limitations concern the extraction of tweet sentiments. Misspelled words are frequent on social media such as Twitter. These words may be skipped over by the sentiment analyzer, which prevents polarized sentiment to be extracted. Subsequent research can resort to a program that detects misspelled words and replace them with their corrected form.

The present experiment is not able to detect repeats of information in Tweets. This affects polarity scores the same positive/negative event appears in multiple publications and is overcounted. Future research can make use of an algorithm that detect and remove repeats, thus obtaining a polarity score that represents the actual media attitude more accurately.

Conclusion

In this paper, the role of polyglotism to Twitter sentiment analysis was examined. Various machine learning models, namely Linear Regression and Random Forest, were utilized. The goal was to determine whether non-English text data have a significant impact on the overall sentiment, as well as the best machine learning model for stock price prediction. Tweets were collected over a month, and polarity is extracted using the Vader lexicon. Non-English text segments are treated by translating into English prior to sentiment analysis. Prediction models are trained using both untranslated and translated tweets. Finally, the models are tested and performance was compared among models.

The results show that stock prediction using Twitter is overall inaccurate. For example, the models predicting the prices of Desjardins exhibited mean square errors between 0.0555 and 0.0787. Taking non-English texts into consideration decreased mean square error by 25.8% in the Linear Regression model, and 13.6% in the Random Forest model. Random forest exhibited slightly superior performance, with mean square error decreased by 18.4% and 5.0% for untranslated and translated models, respectively. The results suggest that non-English tweets have a significant impact on the overall sentiment; hence, developing methods that accurately capture sentiments in polyglot text data can improve the performance of state-of-the-art models.

Acknowledgements

I would like to thank my mentors, Ganesh Mani and Alfred Renaud, for the guidance they provided to successfully complete this research project.

References

Tetlock, Paul C. (2005) Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *Journal of Finance*. https://ssrn.com/abstract=685145

Jin, Z., Yang, Y., & Liu, Y. (2020). Stock closing price prediction based on sentiment analysis and LSTM. *Neural Computing & Applications*, *32*(13), 9713–9729. https://doi-org.ezproxy.marianopolis.edu/10.1007/s00521-019-04504-2

Mendoza-Urdiales, R. A., Núñez-Mora, J. A., Santillán-Salgado, R. J., & Valencia-Herrera, H. (2022). Twitter Sentiment Analysis and Influence on Stock Performance Using Transfer Entropy and EGARCH Methods. *Entropy*, *24*(7), N.PAG. https://doi-org.ezproxy.marianopolis.edu/10.3390/e24070874

Kim, J., Jung, H.-Y., Lee, Y., & Lee, J.-H. (2009). Conveying Subjectivity of a Lexicon of One Language into Another Using a Bilingual Dictionary and a Link Analysis Algorithm. *International Journal of Computer Processing of Languages*, *22*(2/3), 205–218. https://doi-org.ezproxy.marianopolis.edu/10.1142/S1793840609002044

Tellez, E. S., Miranda-Jiménez, S., Graff, M., Moctezuma, D., Suárez, R. R., & Siordia, O. S. (2017). A simple approach to multilingual polarity classification in Twitter. *Pattern Recognition Letters*, *94*, 68–74. https://doi-org.ezproxy.marianopolis.edu/10.1016/j.patrec.2017.05.024

Xiu, D. (2022, November 8). *Expected Returns and Foundation Models of Language*. Seminar presented at the meeting of NYU Courant Mathematics Department.

Statistics Canada. (2011) *Visual Census.* https://www12.statcan.gc.ca/census-recensement/2011/dp-pd/vc-rv/index.cfm?Lang=ENG&VIEW=D&GEOCODE=462&TOPIC_ID=4