Embedding-based dictionary induction

Marina Sedinkina, Nikolas Breitkopf, Hinrich Schtze

Ludwig Maximilian University of Munich
Center for Information and Language Processing

April 20, 2020

Motivation

Finance research: Can the stock market be predicted and explained?

- **Textual analysis** to examine the **sentiment** of:
  - corporate 10-K reports
  - newspaper articles
  - press releases
  - investor message boards

Corporate Report Example

As filed with the Securities and Exchange Commission on October 14, 2015.

Registration No. 333-

UNITED STATES
SECURITIES AND EXCHANGE COMMISSION
Washington, D.C. 20549

FORM S-1
REGISTRATION STATEMENT
Under
The Securities Act of 1933

SQUARE, INC.
(Exact name of Registrant as specified in its charter)

Delaware	7372	80-0429876
(State or other jurisdiction of incorporation or organization)	(Primary Standard Industrial Classification Code Number)	(I.R.S. Employer Identification Number)

1455 Market Street, Suite 600
San Francisco, CA 94103
(415) 375-3176
(Address, including zip code, and telephone number, including area code, of Registrant's principal executive offices)

Jack Dorsey
Chief Executive Officer
Square, Inc.
1455 Market Street, Suite 600
San Francisco, CA 94103
(415) 375-3176
(Name, address, including zip code, and telephone number, including area code, of agent for service)

Copies to:

Steven E. Bochner David J. Segre Tony Jeffries Calise Y. Cheng Wilson Sonsini Goodrich & Rosati, P.C.	David C. Karp Ronald C. Chen Gordon S. Moodie Wachtell, Lipton, Rosen & Katz 51 West 52nd Street	Dana R. Wagner Sydney B. Schaub Tait O. Svenson Square, Inc. 1455 Market Street, Suite 600	William H. Hinman, Jr. Daniel N. Webb Simpson Thacher & Bartlett LLP 2475 Hanover Street

Motivation

- Can **negative word classifications** be effective in **measuring sentiment**?

Motivation

- Can **negative word classifications** be effective in **measuring sentiment**?
- **YES** $\Rightarrow$ sentiment has significant **correlations** with **financial variables** (Antweiler and Frank (2004), Tetlock (2007), Loughran and McDonald (2011)).

Loughran and McDonald (2011)

- **Loughran and McDonald (2011)** construct lists of words, that reflect sentiment in financial contexts.
- 10-K Sample:
  - CRSP (Center for Research in Security Prices) - provider of stock market data
  - 50,115 firm-years of 8,341 unique firms
- Word list construction:
  - **Manually** examine all words in at least 5% of the documents and determine most likely usage in financial documents

Category	**Amount of words**
negative	2,346
positive	354
uncertainty	252
litigious	748
other	457
total	4,157

Loughran and McDonald (2011) - Sample Words

Category	**Most frequently used words**
negative	loss, losses, termination, against, default, closing
positive	effective, benefit, able, gain, greater, good
uncertainty	may, approximately, could, risk, risks, believe
litigious	shall, amended, herein, law, contracts, laws
weak modal	may, could, possible, might, depend, depending

Loughran and McDonald (2011) - Findings

L&M (2011) find a link between sentiment and . . .

- 10-K filing returns
- trading volume
- return volatility
- fraud
- material weakness
- unexpected earnings.

L&M (2011) - Post-Filings Returns

If sentiment matters, firms reports (filings) with a high measure of negative words should experience negative excess returns around the filing date

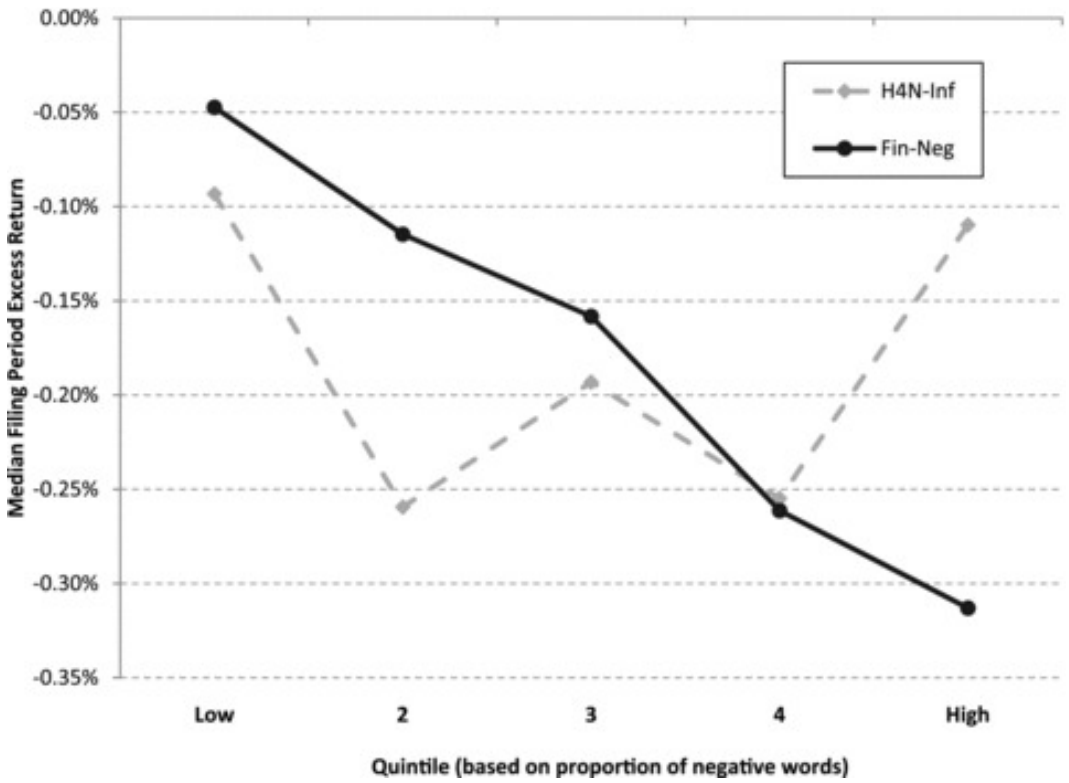


Figure: 4-day excess return around 10-K filing dates

Our Approach

- Use **word2vec** to create word embeddings of the SEC (U.S. Securities and Exchange Commission) 10-K filings corpus
- **SEC 10-K filings**:
  - annual reports: "Business Description", "Company Background", "Risk", "Management", "Legal Issues"
  - Period: 1994 to 2016
  - 60,060 firm-years of 15,360 unique firms
- Use **Support Vector Machines (SVM)** to classify words into sentiment categories
  - L&M (2011) word lists are used as training data
  - SEC words are used as test data

Results of word2vec embedding and SVM classification

Category	**LM**	**NEW**	Most frequently used **new** words
negative	2,346	1,735	untrue, avoid, causes, causing, false
positive	354	1,056	focus, help, expertise, goal, maximize
uncertainty	252	71	estimates, estimate, forecast, subjective
litigious	748	218	trustee, district, validly, registrar

Validation of sentiment over finance variables - Finance Research

t-Statistics of panel regression with both old and new sentiment-related variables (Nikolas Breitkopf)

Dep. Variable	**Dict**	Neg
Excess return	LM	-0.6
	+NEW	-3.53

Significant!

Validation of sentiment over finance variables - NLP

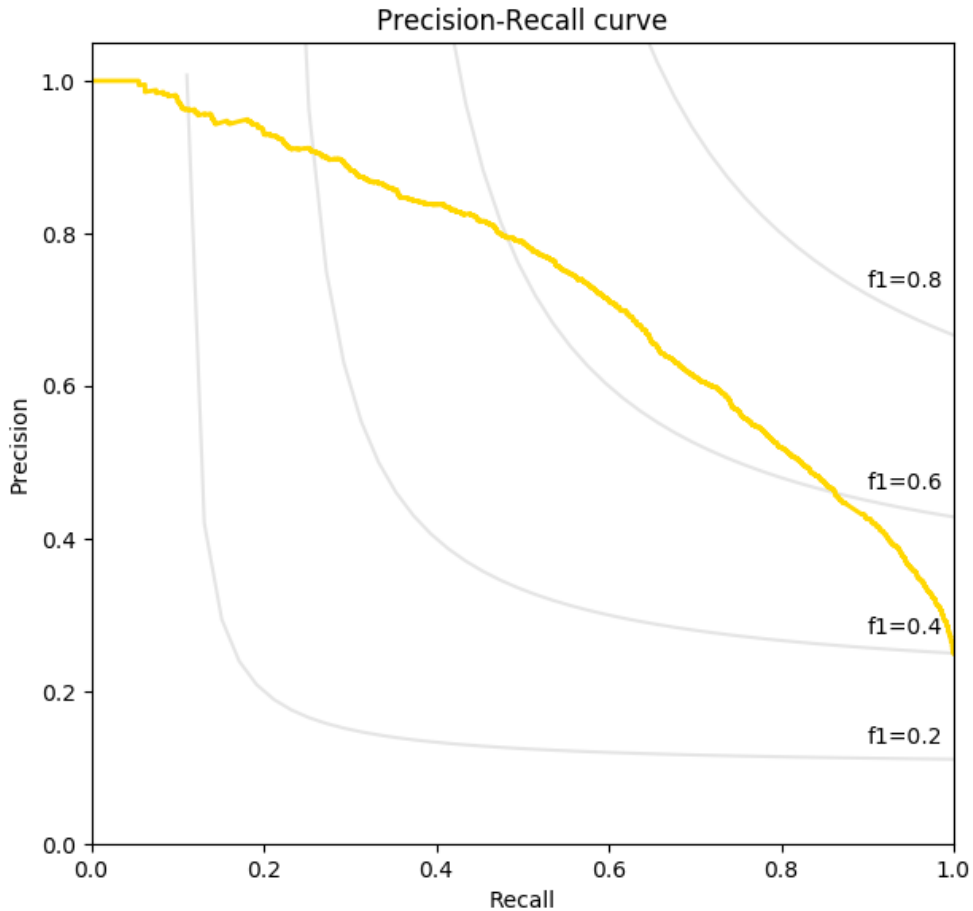
- Finance variable = 3-year excess return
- The sample of 60,060 firm-years is divided into 2 portfolios based on the proportion of 3-year excess return value:
  - Companies that are in the lower quantile ($\leq$ 25%), i.e have negative excess return
  - Companies that are in the upper quantile ($>$ 25%), i.e the rest

Validation of sentiment over finance variables - NLP

- Each sample (document) is represented as a vector of word counts from this document (**"bag of words" method**).
- Later $\Rightarrow$ use only negative word counts from this document.
- Train SVM with 80% of the data
- Test on 20% of the remaining data predicting whether a company is in the lower quantile ($\leq$ 25%).

Results - bag of words

Prediction whether a company is in the lower quantile ($\leq 25\%$) can be made with F1 $= 65\%$



Limitations

Limitations: Training set contains future observations

Possible Solution:

- Perform "rolling" classification, where the test set consists only of future filings:
  - Training set is constructed from observations on the years 2011-2013, Test set are samples from 2014
  - Training set from 2012-2014 and Test 2015
  - ...

Future Goal

- To build an application (sentiment analysis) to explain excess returns
- Can the excess returns be better explained by the new / modified dictionary?