Wisdom of Crowds: Cross-sectional Variation in the Informativeness of Third-Party-Generated Product Information on Twitter

By

Vicki Wei Tang
McDonough School of Business
Georgetown University
Email: wt29@georgetown.edu

July 2017

--Data Description Sheet—
Compliance with Journal of Accounting Research Data Policy

1. A description of which author(s) handled the data and conducted the analyses.

This paper is sole-authored paper and the author handled the data and conducted the analyses herself. Data handling and analyses were performed as described in more detail below (see #2 in particular)

2. A detailed description of how the raw data were obtained or generated, including data sources, the specific date(s) on which data were downloaded or obtained, and the instrument used to generate the data (e.g., for surveys or experiments). We recommend that more than one author is able to vouch for the stated source of the raw data.

The raw data was obtained from both publicly available data sources, including both COMPUSTAT and IBES, and a private data company. Company-level financial information was obtained from COMPUSTAT and analyst forecast information was obtained from the summary tape of I/B/E/S.

The data company, Likefolio.com, provided the first installment of raw data on Twitter comments on 02/15/2016 and the second installment of raw data on Twitter comments on 07/11/2016. For both installments, Likefolio.com first scans all tweets on Twitter and identifies the holder of a given tweet (that is, the person or company who initiates it) and the target of a given tweet (that is, the entity the tweet is discussing). Because the study is interested in the incremental information content of third-party generated comments about a company's products and brands, the target of the selected tweets is limited to products and brands and the holder of the selected tweets is limited to third parties, not

the company itself. Next, Likefolio.com uses a combination of knowledge-based techniques and statistical methods to classify the content of each selected tweet. The first task of content analysis is classifying whether a tweet mentions a recent past purchase of a given product or brand or an intention to do so in the future. The content of each tweet is classified as either "with mentioning of purchase" or "without mentioning of purchase". The second task is classifying the *valence* of a given tweet—whether the opinion expressed is positive, negative, or neutral. Last, Likefolio.com uses its proprietary information to map various products and brands to the businesses that own them and summarizes selected tweets at the *company* level.

The raw data provided by Likefolio.com include four statistics at the company-day level from January 1, 2012 to December 31st, 2015. To control for the growth of Twitter over time, all four statistics are normalized on a per million basis. The first statistic (PRODUCTTWEET) is measured as the number of tweets that discuss a company's products and brands divided by the total number of tweets circulated on Twitter on a given day (in millions). The second statistic (PURCHASETWEET) is measured as the number of tweets that explicitly indicate a recent past purchase of a company's products and brands or an intention to do so in the future divided by the total number of tweets circulated on Twitter on a given day (in millions). The third statistic (POSITIVETWEET) is measured as the number of tweets that convey a positive assessment of a company's products and brands divided by the total number of tweets circulated on Twitter on a given day (in millions). The fourth statistic (NEGATIVETWEET) is measured as the total number of tweets that convey a negative assessment of a company's products and brands divided by the total number of tweets circulated on Twitter on a given day (in millions).

Both the author and the representative from Likefolio.com (his identity is provided to the editor) vouch for the stated sources of the raw data.

3. If the data are obtained from an organization on a proprietary basis, the authors should privately provide the editors with contact information for a representative of the organization who can confirm data were obtained by the authors. The editors would not make this information publicly available. The authors should also provide information to the editors about the data sharing agreement with the organization (e.g., non-disclosure agreements, any restrictions imposed by the organization on the authors, such as restrictions to publish certain results).

The author has provided the editor the contact information for a representative of Likefolio.com. The author has signed a data sharing agreement with the data provider. For contractual reasons, the author is not allowed to share and make available information extracted from Twitter along with the calculation and collection methods with any other person or legal entities. Below the author quotes from the data sharing agreement that imposes such restrictions:

"Title, copyright, intellectual property rights and distribution rights of the Social Data remain exclusively with the Vendor. Intellectual property rights include the look and feel

of the Social Data along with the calculations and collection methods. This Agreement constitutes a license. The Social Data may be loaded only onto computers owned by Georgetown University and its employees, or those directly involved in the research project. The rights and obligations of this Agreement are organization rights granted to the Licensee only. The Licensee may not transfer or assign any of the rights or obligations granted under this Agreement to any other person or legal entity. The Licensee may not make available the Social Data for use by one or more third parties and is not in any way a transfer of ownership rights to the Social Data. Failure to comply with any of the terms under the License section will be considered a material breach of this Agreement."

To demonstrate the nature of the raw data on Twitter comments, the author presents two illustrative entries below:

DATE	*TICKER*	*PRODUCTTWEET*	*PURCHASETWEET*	*POSITIVETWEET*	*NEGATIVETWEET*
XXX	*YYY*	*76*	*3*	*24*	*0*
ZZZ	*UUU*	*54,818*	*12,508*	*8,725*	*816*

4. A complete description of the steps necessary to collect and process the data used in the final analyses reported in the paper. For experimental and survey papers, we require information about the instructions and instruments used to generate the data, subject eligibility and/or selection, as well as any exclusion criteria. The full set of instructions and instruments can be provided in the online appendix.

The author describes data in section 3 of the paper. The author uses IBM SPSS Statistics Version 24 application to process the data used in the final analyses reported in the paper, in which commands are entered from menus in a *graphical* interface and not from *code*. Accordingly, the author provides detailed step-by-step procedures and commands.

Step 1: Download all quarterly analyst forecasts of sales from the I/B/E/S summary tape on 12/09/2016. First use "IDENTIFY DUPLICATE CASES" by official ticker (labeled as OFTIC in I/B/E/S) and the fiscal period end date (labeled as FPEDATS in I/B/E/S) and then use "SELECT CASES" commands to select the last analyst forecast in calendar time (labeled as STATPER in I/B/E/S) for a given company and a given fiscal period end date. Next delete 83 observations where realized sales were missing on the ground that lack of information on realized sales from I/B/E/S makes it impossible to calculate unexpected sales growth in the final analyses. This dataset is named as "*LASTFORECAST.SAV*".

Step 2: Download the COMPUSTAT quarterly universe starting from 01/01/2010 up to the download date of 12/30/2016. Delete those observations where revenue is less than zero for a particular fiscal quarter because those observations do not make economic sense. A very small number of companies have multiple observations for the same fiscal period. The author manually checks those observations and finds that multiple observations for the same fiscal period of a given company are largely identical and the last observation has the most comprehensive amount of data items. Accordingly, if there

are multiple observations for the same fiscal period end date for a given company, the author keeps the last observation.

Step 3: Use the "LAG" command to create the necessary lagged variables in the quarterly COMPUSTAT file in order to calculate both the dependent variable and explanatory variables. For example, in order to calculate the same-quarter sales growth in quarter q (labeled as $SAMEQUARTER_SALESGROWTH_{i,q}$ in the paper), the author first creates a fiscal quarter index by adding up "fiscal year multiplied by four" and "fiscal quarter". Then the author takes the lag in the order of 4 to get sales in the same quarter in the previous fiscal year (sales in quarter q-4). Next the author uses "TRANSFORM" command to calculate the same-quarter sales growth as the difference between sales in quarter q and sales in quarter q-4 divided by sales in quarter q-4.

Step 4: As information on advertising expenses is available only on the annual basis, download the company name, the company identifier (GVKEY), the fiscal year, the annual sales and the annual advertising expense from the COMPUSTAT annual universe starting from 01/01/2010 up to the download date of 12/31/2016. Use the LAG command in SPSS to take the lag in the order of 1 to obtain the annual sales in the most recent fiscal year and the annual advertising expense in the most recent fiscal year, both of which are needed to calculate the pro-rated quarterly advertising expenses.

Step 5: Use the "MERGE FILES" command to merge the quarterly COMPUSTAT file with the annual COMPUSTAT file by both GVKEY and FISCAL YEAR (by checking "match cases by key variables" where key variables are GVKEY and FISCAL YEAR) to form the dataset that contains all financial information on a company-quarter basis. The only complication is the following: as advertising information is available only on annual basis, the author prorates the annual advertising expense by the proportion of sales volume each quarter in order to maintain a consistent quarterly measurement window for both the dependent variable and the explanatory variables. The prorated quarterly advertising expense is measured as the annual advertising expense multiplied by the proportion of quarterly sales over annual sales during the most recent fiscal year. The author then use "TRANSFORM" command to calculate the ratio of the prorated advertising expense over sales during the most recent fiscal quarter and use the calculated ratio as the proxy for the intensity of advertising activities during the most recent fiscal quarter (labeled as $ADVERTISE_{i,q-1}$ in the paper). The merged dataset is named as "*COMPALL.SAV"* and is dated on 01/02/2017.

Step 6: Use the "MERGE FILES" command to merge the dataset extracted from COMPUSTAT (COMPALL.SAV) with the dataset extracted from I/B/E/S (LASTFORECAST.SAV) by both OFFICIAL TICKER and the fiscal period end date (labeled as DATADATE and FPEDATS in COMPUSTAT and I/B/E/S respectively). The merged dataset is named as *"COMP_IBES.SAV*" and is dated on 01/02/2017.

Step 7: Download the proprietary dataset that contains information on Twitter comments on a company-day basis on 01/03/2017. Based on the four numbers provided for each ticker-day combination, the author calculates two statistics, PURCHASE and POSITVE,

on a company-day basis. The first statistic (PURCHASE) is measured as the number of tweets that explicitly indicate a recent past purchase of a company's products and brands or an intention to do so in the future, which is equal to PURCHASETWEET in the raw data. The second statistic (POSITIVE) is measured as the ratio of the number of tweets that convey a positive assessment of a company's products and brands over the combined number of tweets that convey a non-neutral (positive or negative) assessment of a company's products and brands. In particular, the author uses the "TRANSFORM" command to calculate POSITIVE as dividing POSITIVETWEET by the sum of POSITIVETWEET and NEGATIVETWEET. This dataset is named as "TWITTER_RAW.SAV".

Step 8: As the dependent variables, the same-quarter sales growth and unexpected sales growth, are measured on a company-quarter basis, Twitter statistics need to be transformed to the company-quarter level. In particular, on a given date for a particular company, the author uses "CREATE TIME SERIES" command and takes the "MOVING AVERAGE" function to calculate the moving averages of both PURCHASE and POSITIVE over a three-month period that starts from 91 days prior and ends on the given date. The author names the two moving averages as AVG_PURCHASE and AVG_POSITIVE respectively, both of which are the variables of interest in the final analyses. The dataset that contains PURCHASE and POSITIVE averaged over the quarterly interval is named as "TWITTER_AVG.SAV" and is dated on 01/03/2017.

Step 9: Use the "MERGE FILES" command to merge the dataset on TWITTER (TWITTER_AVG.SAV) with the dataset named as COMP_IBES.SAV by both OFFICIAL TICKER and the fiscal period end date. While merging, select those observations from "TWITTER.SAV" where the variable "date" is equal to "the fiscal period end" from "COMP_IBES.SAV" for the same company. The merged dataset is named as "ALL.SAV" and is dated on 01/05/2017. In this dataset, all observations/variables are at the company-quarter level.

Step 10: The private data company provided additional information on Twitter comments by the identity of tweet handles for a subsample of firms that are classified as business-to-consumer firms in the paper on 03/10/2017. The additional information on Twitter comments by tweet handles is provided on a *company-month* basis. As stated in the paper, the identity of tweet handles is classified into three broad categories: the media, product experts and the crowd. The additional information allows the author to calculate three sets of statistics. The first set is the percentage of product tweets initiated by the media, the percentage of product tweets initiated by product experts, and the percentage of product tweets initiated by the crowd. The sum of the three percentages is equal to 100%. The second set is PURCHASE initiated by the media, PURCHASE initiated by product experts, and PURCHASE initiated by the crowd. The third set is POSITIVE initiated by the media, POSITIVE initiated by product experts, and POSITIVE initiated by the crowd. As the raw data on Twitter comments by tweet handles are at the company-month level, the raw data needs to be converted into the company-quarter level in the final analyses. For a given company, for each calendar month, the author uses "CREATE TIME SERIES" command and takes the "MOVING AVERAGE" function to

calculate the moving averages of PURCHASE by tweet handles and POSITIVE by tweet handles for a three-month period that starts from two months prior and ends on the given month. The author labels the moving averages of PURCHASE initiated by the media, product experts and the crowd as MEIDA_AVG_PURCHASE, EXPERT_AVG_PURCHASE and CROWD_AVG_PURCAHSE respectively. The author labels the moving averages of POSITIVE initiated by the media, product experts and the crowd as MEDIA_AVG_POSITIVE, EXPERT_AVG_POSITIVE and CROWD_AVG_POSITIVE respectively. All six variables are variables of interest in the final analyses to product the main results as reported in table 7 and table 9. The author creates a new variable that captures the calendar month index (MONTHINDEX). MONTHINDEX is equal to "calendar year" multiplied by 12 plus "calendar month". The dataset with information on Twitter comments by tweet handles is named as "TWITTERAVG_HANDLE.SAV" and is dated on 03/12/2017.

Step 11: Delete observations with missing value for both the dependent variable and all explanatory variables as specified in model (1) and model (2) from "ALL.SAV". The newly created dataset is named as "COMP_TWITTER.SAV" and is dated on 03/12/2017, which is used to produce the main results as reported in table 5. The subsample of observations for companies whose major customer base are individuals (labeled as the business-to-consumer subsample in the paper) are used to perform the analysis to produce results reported in table 6.

Step 12: Keep the subsample of observations for companies whose major customer base are individuals (labeled as the business-to-consumer subsample in the paper) from "COMP_TWITTER.SAV". Create a calendar month index where MONTHINDEX is equal to taking the "DATEYEAR" function of DATADATE multiplied by 12 plus taking the "DATEMONTH" function of DATADATE. The author uses "MERGE FILES" command to merge the subsample with the dataset named as "TWITTERAVG_HANDLE.SAV" by both OFFICIAL TICKER and MONTHINDEX. While merging, select those observations where MONTHINDEX from "TWITTERAVG_HANDLE.SAV" is equal to MONTHINDEX from the selected subsample of "COMP_IBES.SAV" for the same company. Delete observations with missing value for both the dependent variable and all explanatory variables as specified in model (3). Rank the dependent variable, unexpected sales growth, and explanatory variables by the fiscal period into 200 groups. Deletes three observations where the mean forecast of sales is less than zero because they do not make economic sense. To mitigate the influence of extreme outliers, delete the top 0.5% and the bottom 0.5% of observations. The newly created dataset is named as "IBES_TWITTER.SAV" and is dated on 03/17/2017, which is used to produce main results as reported in table 7, table 8 and table 9.

5. The computer programs or code used to convert the raw data into the final dataset used in the analysis plus a brief description that enables other researchers to use this program. The purpose of this requirement is to facilitate replication and to help other researchers

The author uses SPSS commands that are entered from menus in a graphical interface Statistics to convert raw data to the final dataset used in the analysis. The step-by-step procedures and commands are provided in #4. The author also copies the commands that produce the final results as reported in the tables from the log file onto a syntax file in SPSS. The SPSS file "TANG_TABLE56.SPS" uses "COMP_TWITTER.SAV" as input and yield the content of the main analysis as reported in table 5 and table 6 as output. The SPSS file "TANG_TABLE789.SPS" uses "IBES_TWITTR.SAV" as input and yield the content of the main analysis as reported in table 7, table 8 and table 9 as output. The EXCEL file "GVKEY_TABLE56" lists all the 1840 GVKEYs from the dataset named as "COMP_TWITTER.SAV" that produces results in table 5 and table 6. The Excel file "GVKEY_TABLE789" lists all the 159 GVEKYS from the dataset named as "IBES_TWITTER.SAV" that produces main results as reported in table 7, 8 and 9.

6. An assurance that the data and programs will be maintained by at least one author (usually the corresponding author) for at least six years, consistent with National Science Foundation guidelines.

The author will maintain all data and programs for at least six years.