Online Appendix for "Box Office Buzz: Does Social Media Data Steal the Show from Model Uncertainty When Forecasting for Hollywood?"

Steven Lehrer[†] and Tian Xie[‡]

[†]Queen's University, NYU Shanghai, and NBER, lehrers@queensu.ca
[‡]Xiamen University, xietian001@hotmail.com

November, 2016

Abstract

This is the online appendix regarding *Lehrer and Xie (2016)*. Five sections are included, which provide further details regarding (1) the data utilized, (2) econometric theory, (3) implementation of the relative out of sample prediction efficiency experiment, (4) initial motivation, and (5) additional empirical results.

A Further Details on the Data

In our analysis, we elected to concentrate only on movies whose budgets were between 20 to 100 million dollars. This sample selection criteria was suggested to us by members of IHS film consulting unit and accounts for 41.4% of all releases in the time period studied. We believe it is reasonable since these films receive standard amounts of social media marketing and are destined for national release. There are surprisingly few films with budgets over 100 million dollars (15.4%) and we feel that a different set of candidate models is likely needed for both these big-budget films as well as for many "art-house" films that have small budgets. Thus, while adding more films is possible for our estimation approach, additional computational demands are introduced by increasing the number of candidate models.[1] Our aim is to evaluate whether in situations with remarkably similar entertainment products, are social media and model uncertainty empirically important.

A.1 Sentiment Description

To measure purchasing intentions from the universe of Twitter messages, sentiment specific to a particular film is calculated using an algorithm developed by Janys Analytics for IHS. Specifically, this algorithm which is based on Hannak et al. (2012) involves textual analysis of movie titles and movie key words. In a message that mentions a specific film title or key word, sentiment is calculated by examining the emotion words and icons that are captured in the same Twitter message.[2]

In total, each of 75,065 unique emotion words and icons that appeared in at least 20 tweets between January 1st, 2009 to September 1st, 2009 are given a specific value that is determined using emotional valence.[3] To calculate the sentiment index for the film, a weighted average of the sentiment of the scored words in all of the messages associated with a specific film during a time period is then calculated. This overall sentiment score indicates the propensity for which there is a positive emotion tweet related to that movie.

Since opinions regarding a film likely vary over time with the release of different marketing devices to both build awareness and increase anticipation, IHS film consulting unit suggested to calculate sentiment over different time periods. That is, suppose the movie release date is T, then we separately calculate sentiment in ranges of weeks corresponding to T–21/–27, T–14/–20, T–7/–13, T–4/–6, T–1/T–3. In the DVD analysis, we additionally incorporate measures of sentiment in ranges corresponding to T+0, T+1/+7, T+8/+14, T+15/+21,

[1] We should note that our sample contains few sequels and seasonality appears to play no role unlike prior work evaluating box office revenue such as Einav (2007). In Appendix E.2, we do present some results that utilize this additional information (and continue to find statistical insignificance) but do not utilize this information in the remainder of the main text to gain computational savings.

[2] Twitter messages are capped at 140 characters and often contain acronyms and Twitter specific syntax such as hashtags presenting challenges to using traditional sentiment inference algorithms.

[3] Emotional valence is a term frequently used in psychology that refers to the intrinsic attractiveness (positive valence) or aversiveness (negative valence) of an event, object, or situation.

and T+22/+28.[4] We also include measure corresponding to the volume of Twitter messages for each specific film in these time ranges. For the films analyzed and considering open box office, the volume of Twitter message is 1,100,439; and for DVD, this number is 3,433,413. In general, the day of film release witnesses a significantly higher volume of Tweets than any another day.

We next present the empirical probability distribution function of sentiment data for each time interval using kernel-based method in Figures A1 and A2. Due to design of the emotional valence algorithm, the mean of sentiment variables is around 75, but does exhibit significant variation in each time interval. We should stress that there has been substantial evaluation of the sentiment inference algorithm developed by Janys Analytics for IHS. Hannak et al. (2012) compare this sentiment inference methodology score to one calculated by users of Amazon Mechanical Turk and find that they are strongly positively correlated with $\rho = 0.6525$. An additional advantage is that the sentiment inference algorithm is easy to regularly update to readjust the frequency at which a specific work is associated with a positive emotion in calculating the initial values that enter the sentiment calculator to adjust to potential changes in the Twitter user population.

A final important issue to briefly discuss is that the demographics of Twitter users differs markedly from the national population. Mislove et al. (2011) document that these users are predominately male and located in urban areas, but point out these calculations are based on self-reported profiles. Further, these authors note that the male bias is declining rapidly. Despite the self-selection of these users, we believe that this sample of users is likely highly correlated with the characteristics of moviegoers and DVD purchasers so is relevant to study. After all, research in marketing indicates that everyday consumers often seek like-minded amateurs' opinions (for example, Chakravarty, Liu, and Mazumdar (2009) and Holbrook (1999)).

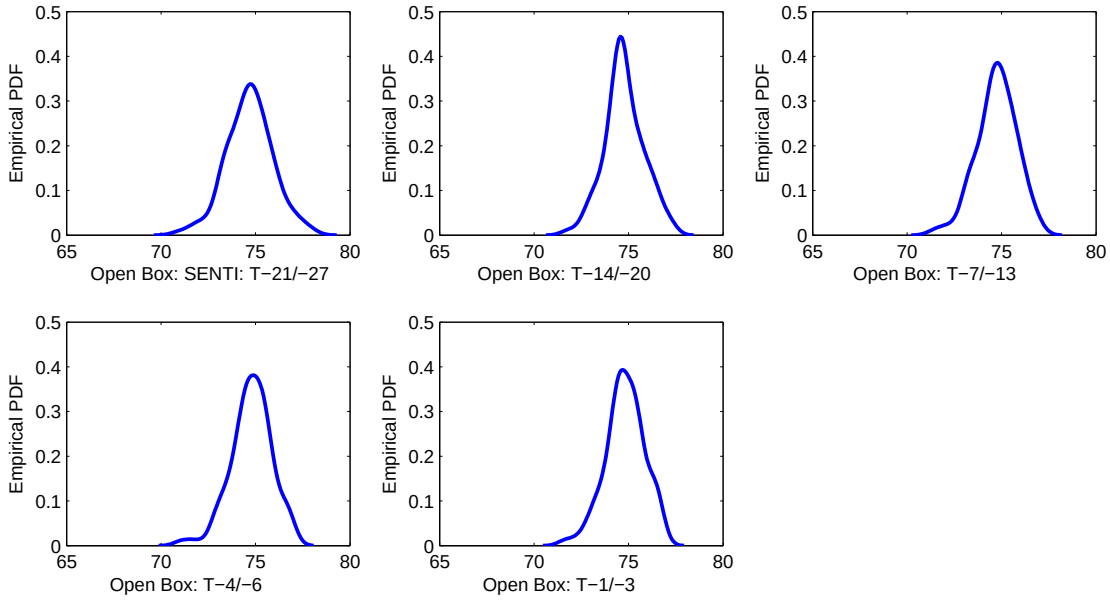
B Further Motivation for Model Uncertainty with Social Media Data in Empirical Economics

While theory may not directly guide our choice of model specification it does provide a clear rationale for using least squares model averaging. Specifically, both accounting for social media data and allowing for model uncertainty is consistent with a growing body of work in behavioral economics.[5] Whereas, the standard neoclassical economic model assumes that people care only about maximizing their own payoffs, a large body of research has

[4]We use the sentiment data before the release date in equations that forecast the opening weekend box office. After all, reverse causality issues would exist if we include sentiment data after the release date. These latter measures are only considered in equations that forecast DVD and Blu-Ray sales.

[5]Research in behavioral economics focuses on identifying these forms of biases and exploring their implications for standard economic models. That said, the examples that we discuss in the remainder of the paragraph do not consider the network structure of Twitter. It is likely that there is what psychologists term as belief polarization and research in computer science has shown that individuals give more weight to messages from those that are considered strong ties relative to weak ties.

Figure A1: Empirical PDF of Sentiment Variables for Open Box Office



shown that social forces also play a large role in decision making.[6] Further, an economic theory literature has developed analyzing (near) costless communication in a sender-receiver game framework. In an important paper, Crawford (2003) demonstrates if there is the possibility of interacting with boundedly rational receiver in this setting, multiple equilibria arise, including those in which the sender lies in pre-play communication of their intentions.[7] Thus, there are many strategic rationales underlying the intentions' one communicates to their followers via messages on the social web.[8]

Indeed there has been a number of economists that have also begun to include social media data in their analysis, Antenucci, Cafarella, Levenstein, Ré, and Shapiro (2014) and Toole, Lin, Muehlegger, Shoag, González, and Lazer (2015) each illustrate the potential of data from the social web to measure economic indicators of labor market activity.[9] Outside of economics, there also has been tremendous growth in academic circles in using data extracted from social media (Facebook, Twitter, Google+, etc) to analyze the economy with these tools. For example, Mishne and Glance (2006) proposed using Blogger sentiment to predict movie sales. Bollen, Mao, and Zheng (2011) demonstrated that Twitter mood can

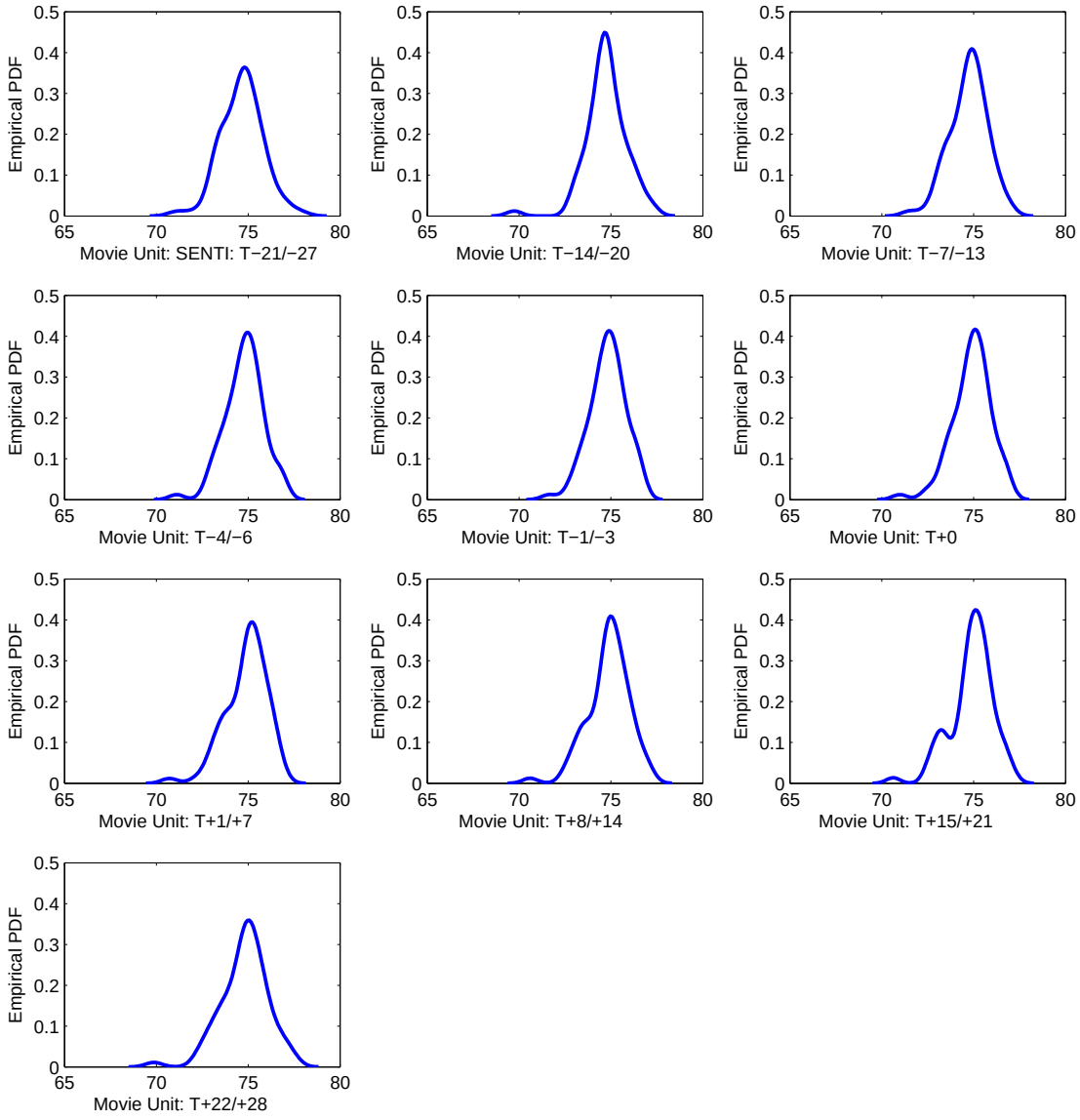
[6]For example, see Lee (2008), Hsu and Lin (2008), and Kleef, Dreu, and Manstead (2010) for details.

[7]In addition, recent experimental research (Andreoni and Bernheim (2009) and Embrey, Fréchette, and Lehrer (2015)) has shown that individuals are more likely to deviate from social norms when they have a means to hide their true intentions.

[8]Further, it is plausible that an individual recipient may interpret an identical message differently based on who the poster is. While the identical message would receive the same individual score, it could, for example, if made by a right wing individual be interpreted as being supportive, whereas it might be deemed as being sarcastic if made by an individual on the other side of the political spectrum.

[9]Einav and Levin (2014) summarize the opportunities and challenges that confront economists wishing to take advantage of large new data sets either obtained from the social web or administrative records.

Figure A2: Empirical PDF of Sentiment Variables for DVD Unit Sales



predict the stock market. In a recent paper, Karabulut (2013) showed that the stock market activity can also be predicted by measures extracted from Facebook messages. Yet, to the best of our knowledge, no prior study considered applying model averaging when forecasting outcomes with explanatory variables extracted from the social web.

C More Details on the Econometric Theory

C.1 Theorem 1

In this subsection, we lay out all the necessary details to derive Theorem 1 in the main text, which demonstrates that if we group regressors into sets of four or larger, the PMA estimator always yields smaller asymptotic risk than the unrestricted least-squares estimator.

We begin by continuing to assume that the DGP follows (3). The regressor $\boldsymbol{X}$ can be partitioned into ordered groups as $\boldsymbol{X} = [\boldsymbol{X}_1, ..., \boldsymbol{X}_m, ..., \boldsymbol{X}_M]$, where $\boldsymbol{X}_m$ is $n \times k_m$ and the total number of regressors is $k = k_1 + ... + k_M$. In this set-up, all models are nested in sequence. Therefore, M groups of regressors generate M potential models. The m^{th} model includes the regressors $\boldsymbol{X}^{(m)} = [\boldsymbol{X}_1, ..., \boldsymbol{X}_m]$ and excludes the remaining regressors. Note that the m^{th} potential model has $k^{(m)} = k_1 + ... + k_m$ regressors and the regressor $\boldsymbol{X}_1$ are included in all models.

The unconstrained least-squares estimator of $\boldsymbol{\beta}$ in the full model is

$$\hat{\boldsymbol{\beta}}_{LS} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \tag{A1}$$

with residual $\hat{\boldsymbol{u}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{LS}$. If we define a $k \times k^{(m)}$ matrix

$$\boldsymbol{S}^{(m)} = \begin{bmatrix} \boldsymbol{I}^{(m)} \\ \boldsymbol{0} \end{bmatrix}$$

where $\boldsymbol{I}^{(m)}$ is an identity matrix with rank $k^{(m)}$, for the m^{th} model, we have $\boldsymbol{X}^{(m)} = \boldsymbol{X}\boldsymbol{S}^{(m)}$. Given the least squares estimator $\hat{\boldsymbol{\beta}}^{(m)} = \left(\boldsymbol{X}^{(m)\top} \boldsymbol{X}^{(m)}\right)^{-1} \boldsymbol{X}^{(m)\top} \boldsymbol{y}$ is $k^{(m)} \times 1$, we can expand it to $k \times 1$ using the $\boldsymbol{S}^{(m)}$ matrix:

$$\tilde{\boldsymbol{\beta}}^{(m)} = \boldsymbol{S}^{(m)} \hat{\boldsymbol{\beta}}^{(m)} = \boldsymbol{S}^{(m)} \left(\boldsymbol{S}^{(m)\top} \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{S}^{(m)}\right)^{-1} \boldsymbol{S}^{(m)\top} \boldsymbol{X}^\top \boldsymbol{y}.$$

The corresponding residual is $\hat{\boldsymbol{u}}^{(m)} = \boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}^{(m)}$. Note that since model M contains all regressors, we have $\hat{\boldsymbol{\beta}}^{(M)} = \hat{\boldsymbol{\beta}}_{LS}$ and $\hat{\boldsymbol{u}}^{(M)} = \hat{\boldsymbol{u}}$.

An averaging estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}(\boldsymbol{w}) = \sum_{m=1}^{M} w^{(m)} \tilde{\boldsymbol{\beta}}^{(m)}.$$

The residual from the averaging estimator is

$$\hat{\boldsymbol{u}}(\boldsymbol{w}) = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\boldsymbol{w}) = \sum_{m=1}^{M} w^{(m)} \hat{\boldsymbol{u}}^{(m)}.$$

We can rewrite PMA in (6) as

$$\text{PMA}_n(\boldsymbol{w}) = \hat{\boldsymbol{u}}(\boldsymbol{w})^\top \hat{\boldsymbol{u}}(\boldsymbol{w}) + 2\sum_{m=1}^{M} w^{(m)} T_m(\boldsymbol{w}), \tag{A2}$$

where $T_m(\boldsymbol{w}) = \hat{\sigma}(\boldsymbol{w})k^{(m)}$ and

$$\hat{\sigma}^2(\boldsymbol{w}) = \frac{\hat{\boldsymbol{u}}(\boldsymbol{w})^\top \hat{\boldsymbol{u}}(\boldsymbol{w})}{n - \sum_{m=1}^{M} w^{(m)} k^{(m)}}$$

Given the estimated weighting vector $\hat{\boldsymbol{w}}$, we define the estimated averaged estimator as

$$\hat{\boldsymbol{\beta}}_A = \sum_{m=1}^{M} \hat{w}^{(m)} \tilde{\boldsymbol{\beta}}^{(m)}. \tag{A3}$$

For convenience, we consider the cumulative weight version of PMA. Let $w_*^{(m)} = w^{(1)} + ... + w^{(m)}$ and define the cumulative weight vector $\boldsymbol{w}^* = \left[w_*^{(1)}, ..., w_*^{(M)}\right]^\top$. Similarly, the estimated cumulative weight vector is $\hat{\boldsymbol{w}}^* = \left[\hat{w}_*^{(1)}, ..., \hat{w}_*^{(M)}\right]^\top$. Note that $\boldsymbol{w} \in \boldsymbol{H_M}$ is equivalent to $\boldsymbol{w}^* \in \boldsymbol{H_M^*}$, where $\boldsymbol{H_M^*} = \left\{\boldsymbol{w}^* : 0 \le w_*^{(1)} \le \cdots \le w_*^{(M)} \le 1\right\}$. Using the cumulative weight vector, the averaged estimator can be written as

$$\hat{\boldsymbol{\beta}}(\boldsymbol{w}) = \hat{\boldsymbol{\beta}}_{LS} - \sum_{m=1}^{M-1} w_*^{(m)} \left(\tilde{\boldsymbol{\beta}}^{(m+1)} - \tilde{\boldsymbol{\beta}}^{(m)}\right)$$

and the estimated averaged estimator is simply $\hat{\boldsymbol{\beta}}_A = \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{w}})$.

Define $t_{m+1}(\boldsymbol{w}) \equiv T_{m+1}(\boldsymbol{w}) - T_m(\boldsymbol{w})$ and $L_m \equiv \left(\hat{\boldsymbol{u}}^{(m)}\right)^\top \hat{\boldsymbol{u}}^{(m)}$, we have

Lemma 1 *The PMA defined in (A2) can be rewritten as*

$$PMA_n(\boldsymbol{w}) = PMA_n^*(\boldsymbol{w}^*) + L_M,$$

where

$$PMA_n^*(\boldsymbol{w}^*) = \sum_{m=1}^{M-1} \left(\left(w_*^{(m)}\right)^2 (L_m - L_{m+1}) - 2w_*^{(m)} t_{m+1}(\boldsymbol{w})\right) + 2T_M(\boldsymbol{w}). \tag{A4}$$

Hence,

$$\hat{\boldsymbol{w}}^* = \underset{\boldsymbol{w}^* \in \boldsymbol{H_M^*}}{\arg\min}\ PMA_n^*(\boldsymbol{w}^*). \tag{A5}$$

Lemma 1 states that the original PMA defined in (A2) can be transformed into $\text{PMA}_n^*(\boldsymbol{w}^*)$ that incorporates the cumulative weight vector $\boldsymbol{w}^*$. Also, $\boldsymbol{w}^*$ can be estimated through convex optimization of $\text{PMA}_n^*(\boldsymbol{w}^*)$.

Note that although the cumulative weight criterion PMA_n^* includes both cumulative weights $\boldsymbol{w}^*$ and model weights $\boldsymbol{w}$ in (A4), $\boldsymbol{w}$ can be linearly transformed into $\boldsymbol{w}^*$ easily (and *vice versa*). We keep both weighting vectors in (A4) for convenience in proofs. Our cumulative weight criterion PMA_n^* does not belong to the class of least squares model averaging criteria defined in Lemma 1 of Hansen (2014). Therefore, results from Hansen (2014) can not be applied to our PMA estimator directly.

We impose the following assumptions.

Assumption 1 *Given a* $1 \times k$ *row vector* $\boldsymbol{x}_i$ *that is a row in* $\boldsymbol{X}$*, we let* $\boldsymbol{Q} = \mathbb{E}(\boldsymbol{x}_i^\top \boldsymbol{x}_i) > \boldsymbol{0}$.

Assumption 2 *There exist some fixed integer* $N < \infty$*, such that*

$$\mathbb{E}[|u_i|^{4(N+1)}|\boldsymbol{x}_i] \leq \kappa < \infty.$$

Assumption 3 *As* $n \to \infty$*,* $n^{1/2}\boldsymbol{\beta}^{(m)} \to \boldsymbol{\delta}^{(m)}$ *for* $m = 2, \cdots, M$.

Assumption 4 *As* $n \to \infty$*,* $k_M/n \to 0$.

Assumptions 1 and 2 are bounding conditions on $\boldsymbol{x}_i$ and u_i. Assumption 3 is a local asymptotic framework and it allows the coefficient $\boldsymbol{\beta}^{(m)}$ to be in a local $n^{-1/2}$ neighbourhood around 0. Note that $\boldsymbol{\beta}^{(1)}$ is not constrained by this framework since it is included in all models. Assumption 4 guarantees that k_M is always smaller than n.

Lemma 2 *Let Assumptions 1–4 hold, as* $n \to \infty$*, we have* $T_m(\boldsymbol{w}) \xrightarrow{p} T_m^0$ *for* $m = 1, \cdots, M$*, where* $T_m^0 = \sigma^2 k^{(m)}$.

Lemma 2 states that $T_m(\boldsymbol{w})$ converges in probability to a term T_m^0 which is independent of $\boldsymbol{w}$ asymptotically. Therefore, combining Lemma 1 with Lemma 2, we can conclude that the term $2T_M(\boldsymbol{w})$ in $\text{PMA}_n^*(\boldsymbol{w}^*)$ can be ignored when estimating $\hat{\boldsymbol{w}}^*$ asymptotically.

Two more assumptions are needed

Assumption 5 $\{y_i, \boldsymbol{x}_i\}$ *is independent and identically distributed with finite fourth moments.*

Assumption 6 *We let* $k_m \geq 4$ *for all* $m > 1$*, which means the regressors* $\boldsymbol{X}_m$ *are grouped into sets of four or larger, except* $\boldsymbol{X}_1$.

Following Hansen (2014), we define the asymptotic risk as

$$R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \lim_{\zeta \to \infty} \liminf_{n \to \infty} \mathbb{E} \min \left\{ n(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \boldsymbol{Q} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}), \zeta \right\},$$

where $\hat{\boldsymbol{\beta}}$ can be any estimated $\boldsymbol{\beta}$. With the above Assumptions and Lemmas, we represent the following theorem that also appears in the main text.

Theorem 1 *Let Assumptions 1 – 6 hold. We have*

$$R(\hat{\boldsymbol{\beta}}_A, \boldsymbol{\beta}) < R(\hat{\boldsymbol{\beta}}_{LS}, \boldsymbol{\beta}), \tag{A6}$$

where $\hat{\boldsymbol{\beta}}_{LS}$ and $\hat{\boldsymbol{\beta}}_A$ are defined in (A1) and (A3) respectively.

Theorem 1 can be seen as a supplement of Theorem 3 in Hansen (2014) and as such we only consider the homoskedastic case. We leave the heteroskedastic case for future research.

C.2 Proof

Proof of Lemma 1 Note that $L_j \geq L_{j+1}$ and $\left(\hat{\boldsymbol{u}}^{(j)}\right)^\top \hat{\boldsymbol{u}}^{(m)} = L_{\max(j,m)}$ by the properties of the least square residuals of this nested model situation. The PMA criterion is then

$$\mathrm{PMA}_n(\boldsymbol{w}) = \sum_{m=1}^{M}\sum_{m=1}^{M} w^{(j)} w^{(m)} L_{\max(j,m)} + 2\sum_{m=1}^{M} w^{(m)} T_m(\boldsymbol{w}). \tag{A7}$$

The first term in (A7) can be rewritten as

$$\begin{aligned}
& (w^{(1)})^2 L_1 + \left((w^{(2)})^2 + 2w^{(2)}w^{(1)}\right) L_2 + \cdots + \left((w^{(M)})^2 + 2(w^{(M)})(w^{(1)} + \cdots + w^{(M-1)}\right) L_M \\
= \; & (w^{(1)})^2 L_1 + \left((w^{(1)} + w^{(2)})^2 - (w^{(1)})^2\right) L_2 + \cdots \\
& \qquad + \left((w^{(1)} + \cdots + w^{(M)})^2 - (w^{(1)} + \cdots + w^{(M-1)})^2\right) L_M \\
= \; & \sum_{m=1}^{M-1} \left(w_*^{(m)}\right)^2 (L_m - L_{m+1}) + L_M.
\end{aligned} \tag{A8}$$

The second term in (A7) is

$$\begin{aligned}
2\sum_{m=1}^{M} w^{(m)} T_m(\boldsymbol{w}) &= 2w^{(1)}(T_1(\boldsymbol{w}) - T_2(\boldsymbol{w})) + 2(w^{(1)} + w^{(2)})(T_2(\boldsymbol{w}) - T_3(\boldsymbol{w})) + \cdots \\
&\quad + 2(w^{(1)} + \cdots + w^{(M-1)})(T_{M-1}(\boldsymbol{w}) - T_M(\boldsymbol{w})) + w_*^{(M)} T_M(\boldsymbol{w}) \\
&= -2\sum_{m=1}^{M-1} w_*^{(m)} (T_{m+1}(\boldsymbol{w}) - T_m(\boldsymbol{w})) + 2T_M(\boldsymbol{w}).
\end{aligned} \tag{A9}$$

Plug (A8) and (A9) in (A7), we get

$$\mathrm{PMA}_n^*(\boldsymbol{w}^*) = \sum_{m=1}^{M-1} \left(\left(w_*^{(m)}\right)^2 (L_m - L_{m+1}) - 2w_*^{(m)} t_{m+1}(\boldsymbol{w}) \right) + 2T_M(\boldsymbol{w}).$$

Moreover, since L_M is unrelated with $\boldsymbol{w}$, we have $\hat{\boldsymbol{w}}^* = \underset{\boldsymbol{w}^* \in \boldsymbol{H}_M^*}{\arg\min}\, \mathrm{PMA}_n^*(\boldsymbol{w}^*)$. ∎

Proof of Lemma 2 A sufficient and necessary condition to prove Lemma 2 is

$$\hat{\sigma}^2(\boldsymbol{w}) = \frac{\hat{\boldsymbol{u}}(\boldsymbol{w})^\top \hat{\boldsymbol{u}}(\boldsymbol{w})}{n - k(\boldsymbol{w})} \xrightarrow{p} \sigma^2, \tag{A10}$$

which implies that $\hat{\sigma}^2(\boldsymbol{w})$ is a consistent estimator of σ^2.

Using equations (A7) and (A8) in the proof of Lemma 1, we obtain

$$\hat{\boldsymbol{u}}(\boldsymbol{w})^\top \hat{\boldsymbol{u}}(\boldsymbol{w}) = \sum_{m=1}^{M-1} \left(w_*^{(m)}\right)^2 (L_m - L_{m+1}) + L_M,$$

Given $w_m^* \leq 1$ for any $m = 1, \cdots, M$ and Assumption 4, a sufficient condition for (A10) to hold is

$$\frac{L_M}{n} \xrightarrow{p} \sigma^2 \quad \text{and} \quad \frac{L_m - L_{m+1}}{n} \xrightarrow{p} 0$$

for $m = 1, \cdots, M-1$, which can be achieved if

$$\frac{L_m}{n} \xrightarrow{p} \sigma^2 \quad \text{for } m = 1, \cdots, M. \tag{A11}$$

For each m, we have

$$\hat{\boldsymbol{u}}^{(m)} = \boldsymbol{y} - \boldsymbol{X}^{(m)} \hat{\boldsymbol{\beta}}^{(m)} = \left(\boldsymbol{I} - \boldsymbol{P}^{(m)}\right)\boldsymbol{u} + \left(\boldsymbol{I} - \boldsymbol{P}^{(m)}\right)\boldsymbol{X}_-^{(m)} \boldsymbol{\beta}_-^{(m)}$$

where $\boldsymbol{P}^{(m)}$ stands for the projection matrix of $\boldsymbol{X}^{(m)}$, $\boldsymbol{X}_-^{(m)} = [\boldsymbol{X}_{m+1}, \cdots, \boldsymbol{X}_M]$ represents the factors not included in $\boldsymbol{X}^{(m)}$, and $\boldsymbol{\beta}_-^{(m)} = \left[\boldsymbol{\beta}_{m+1}^\top, \cdots, \boldsymbol{\beta}_M^\top\right]^\top$ is the associated coefficient vector of $\boldsymbol{X}_-^{(m)}$.

Then, we have

$$\frac{L_m}{n} = \frac{1}{n}\boldsymbol{u}^\top \left(\boldsymbol{I} - \boldsymbol{P}^{(m)}\right)\boldsymbol{u} + \frac{1}{n}\left(\boldsymbol{X}_-^{(m)}\boldsymbol{\beta}_-^{(m)}\right)^\top \left(\boldsymbol{I} - \boldsymbol{P}^{(m)}\right)\boldsymbol{X}_-^{(m)}\boldsymbol{\beta}_-^{(m)} + \frac{2}{n}\boldsymbol{u}^\top \left(\boldsymbol{I} - \boldsymbol{P}^{(m)}\right)\boldsymbol{X}_-^{(m)}\boldsymbol{\beta}_-^{(m)} \tag{A12}$$

By a straightforward application of Theorem 2 in Hansen (2007), we have

$$\frac{1}{n}\boldsymbol{u}^\top \left(\boldsymbol{I} - \boldsymbol{P}^{(m)}\right)\boldsymbol{u} \xrightarrow{p} \sigma^2.$$

Also, the third term of (A12) is $o_p(1)$ by the condition $\mathbb{E}(\boldsymbol{u}|\boldsymbol{X}) = \boldsymbol{0}$. By $\boldsymbol{Q} = \mathbb{E}[\boldsymbol{x}_i^\top \boldsymbol{x}_i] > 0$, we have

$$\frac{1}{n}\left(\boldsymbol{X}_-^{(m)}\right)^\top \left(\boldsymbol{I} - \boldsymbol{P}^{(m)}\right)\boldsymbol{X}_-^{(m)} \xrightarrow{p} \left(\boldsymbol{Q}_-^{(m)}\right)^{1/2} \left(\boldsymbol{I} - \boldsymbol{P}^{(m)}\right)\left(\boldsymbol{Q}_-^{(m)}\right)^{1/2}$$

which is finite by Assumption 1 and $\boldsymbol{Q}_-^{(m)} = \frac{1}{n}\mathbb{E}\left(\left(\boldsymbol{X}_-^{(m)}\right)^\top \boldsymbol{X}_-^{(m)}\right)$. By Assumption 2, we have $\boldsymbol{\beta}_-^{(m)} = O(n^{1/2})$. Therefore, the second term of (A12)

$$\frac{1}{n}\left(\boldsymbol{X}_-^{(m)}\boldsymbol{\beta}_-^{(m)}\right)^\top \left(\boldsymbol{I} - \boldsymbol{P}^{(m)}\right)\boldsymbol{X}_-^{(m)}\boldsymbol{\beta}_-^{(m)} \to 0$$

Therefore, condition (A11) is achieved. ∎

Proof of Theorem 1 Theorem 1, Theorem 2, and Theorem 3 of Hansen (2014) established result (A6) for a broad class of linear estimators. Many parts of Hansen's (2014) proof can be applied to our case except for a key procedure. In the proof of Theorem 1, Hansen (2014) showed that

$$\hat{\boldsymbol{w}}^* \xrightarrow{d} \boldsymbol{w}^*(\boldsymbol{Z}+\boldsymbol{\delta}),$$

where $\sqrt{n}(\hat{\boldsymbol{\beta}}_{LS}-\boldsymbol{\beta}) \xrightarrow{d} \mathrm{N}(\mathbf{0},\boldsymbol{V}) \equiv \boldsymbol{Z}$, $\sqrt{n}\boldsymbol{\beta} \to \boldsymbol{\delta}$. And $\boldsymbol{w}^*(\boldsymbol{Z}+\boldsymbol{\delta}) = \underset{\boldsymbol{w}^* \in \boldsymbol{H}_{\boldsymbol{M}}^*}{\arg\min}\, C^*(\boldsymbol{w}^*, \boldsymbol{Z}+\boldsymbol{\delta})$, where

$$C^*(\boldsymbol{w}^*, \boldsymbol{Z}+\boldsymbol{\delta}) = \sum_{m=1}^{M-1}\left(w_m^{*2}(\boldsymbol{Z}+\boldsymbol{\delta})^\top \boldsymbol{Q}\big(\boldsymbol{P}_{\boldsymbol{Q}}^{(m+1)} - \boldsymbol{P}_{\boldsymbol{Q}}^{(m)}\big)\boldsymbol{Q}(\boldsymbol{Z}+\boldsymbol{\delta}) - 2w_m^* t_{m+1}^0\right)$$

with $t_{m+1}^0 = T_{m+1}^0 - T_m^0$, $T_m^0 = \sigma^2 k^{(m)}$, and $\boldsymbol{P}_{\boldsymbol{Q}}^{(m)} \equiv \boldsymbol{S}^{(m)}({\boldsymbol{S}^{(m)}}^\top \boldsymbol{Q}\boldsymbol{S}^{(m)})^{-1}{\boldsymbol{S}^{(m)}}^\top$.

For PMA case, it is straightforward to demonstrate that

$$\begin{aligned}
\sqrt{n}\tilde{\boldsymbol{\beta}}^{(m)} &= \boldsymbol{S}^{(m)}\left({\boldsymbol{S}^{(m)}}^\top\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)\boldsymbol{S}^{(m)}\right)^{-1}{\boldsymbol{S}^{(m)}}^\top\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)\sqrt{n}\hat{\boldsymbol{\beta}}_{LS} \\
&\xrightarrow{d} \boldsymbol{S}^{(m)}({\boldsymbol{S}^{(m)}}^\top\boldsymbol{Q}\boldsymbol{S}^{(m)})^{-1}{\boldsymbol{S}^{(m)}}^\top\boldsymbol{Q}(\boldsymbol{Z}+\boldsymbol{\delta}) \\
&= \boldsymbol{P}_{\boldsymbol{Q}}^{(m)}\boldsymbol{Q}(\boldsymbol{Z}+\boldsymbol{\delta})
\end{aligned}$$

Therefore,

$$\sqrt{n}\big(\tilde{\boldsymbol{\beta}}^{(m+1)} - \tilde{\boldsymbol{\beta}}^{(m)}\big) \xrightarrow{d} \big(\boldsymbol{P}_{\boldsymbol{Q}}^{(m+1)} - \boldsymbol{P}_{\boldsymbol{Q}}^{(m)}\big)\boldsymbol{Q}(\boldsymbol{Z}+\boldsymbol{\delta}). \tag{A13}$$

Similarly, we have

$$\begin{aligned}
L_m - L_{m+1} &= \left(\hat{\boldsymbol{u}}^{(m)}\right)^\top\hat{\boldsymbol{u}}^{(m)} - \left(\hat{\boldsymbol{u}}^{(m+1)}\right)^\top\hat{\boldsymbol{u}}^{(m+1)} \\
&= \left(\hat{\boldsymbol{u}}^{(m)} - \hat{\boldsymbol{u}}^{(m+1)}\right)^\top\left(\hat{\boldsymbol{u}}^{(m)} - \hat{\boldsymbol{u}}^{(m+1)}\right) \\
&= \left(\hat{\boldsymbol{\beta}}^{(m+1)} - \hat{\boldsymbol{\beta}}^{(m)}\right)^\top\boldsymbol{X}^\top\boldsymbol{X}\left(\hat{\boldsymbol{\beta}}^{(m+1)} - \hat{\boldsymbol{\beta}}^{(m)}\right) \\
&= \sqrt{n}\left(\hat{\boldsymbol{\beta}}^{(m+1)} - \hat{\boldsymbol{\beta}}^{(m)}\right)^\top\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)\sqrt{n}\left(\hat{\boldsymbol{\beta}}^{(m+1)} - \hat{\boldsymbol{\beta}}^{(m)}\right)
\end{aligned} \tag{A14}$$

Substitute (A13) into (A14), we have

$$\begin{aligned}
L_m - L_{m+1} &\xrightarrow{d} (\boldsymbol{Z}+\boldsymbol{\delta})^\top\boldsymbol{Q}\big(\boldsymbol{P}_{\boldsymbol{Q}}^{(m+1)} - \boldsymbol{P}_{\boldsymbol{Q}}^{(m)}\big)\boldsymbol{Q}\big(\boldsymbol{P}_{\boldsymbol{Q}}^{(m+1)} - \boldsymbol{P}_{\boldsymbol{Q}}^{(m)}\big)\boldsymbol{Q}(\boldsymbol{Z}+\boldsymbol{\delta}) \\
&= (\boldsymbol{Z}+\boldsymbol{\delta})^\top\boldsymbol{Q}\big(\boldsymbol{P}_{\boldsymbol{Q}}^{(m+1)} - \boldsymbol{P}_{\boldsymbol{Q}}^{(m)}\big)\boldsymbol{Q}(\boldsymbol{Z}+\boldsymbol{\delta}).
\end{aligned} \tag{A15}$$

The equality in (A15) is an application of Lemma 3 of Hansen (2014). Therefore, by (A15),

our Lemma 2, and Assumptions 1 and 5, we find that

$$\begin{aligned} \mathrm{PMA}_n^*(\boldsymbol{w}^*) &= \sum_{m=1}^{M-1} \left(\left(w_*^{(m)}\right)^2 (L_m - L_{m+1}) - 2w_*^{(m)} t_{m+1}(\boldsymbol{w}) \right) + 2T_M(\boldsymbol{w}) \\ &\xrightarrow{d} C^*(\boldsymbol{w}^*, \boldsymbol{Z} + \boldsymbol{\delta}) + 2T_M. \end{aligned}$$

Since (A5) is a convex minimization problem, and $2T_M$ is unrelated with $\boldsymbol{w}$, we can apply the argument of Kim and Pollard (1990) and deduce that

$$\hat{\boldsymbol{w}}^* \xrightarrow{d} \boldsymbol{w}^*(\boldsymbol{Z} + \boldsymbol{\delta})$$

for our PMA estimator. The rest of proofs follows Hansen (2014). ∎

D Further Details on the Relative Out-of-sample Prediction Efficiency Experiment

D.1 Grouping Methods

In the main text, we supplement Theorem 3 in Hansen (2014),[10] allowing this finding to be applied to a broader set of least squares model averaging estimators including the PMA estimator. Since both our Theorem 1 and Theorem 3 in Hansen (2014) require regressors to be grouped into sets of four or larger, we group regressors based on either economic intuition (g_1) or statistical logic ($g2$) as outlined below.

Economic Intuition: We follow our personal economic intuition and placed variables that capture similar characteristics into a single group. Note that we have one group of regressors, $\boldsymbol{X}_1$, that must be included in all models and the number of regressors of $\boldsymbol{X}_1$ can be any number. For open box office, we have

$\boldsymbol{X}_1$: Key variables, Constant, Animation, Family, Weeks, Screens, VOL: T-1/-3
$\boldsymbol{X}_2$: Volume, T-21/-27, T-14/-20, T-7/-13, T-4/-6
$\boldsymbol{X}_3$: Sentiment, T-21/-27, T-14/-20, T-7/-13, T-4/-6, T-1/-3
$\boldsymbol{X}_4$: Rating, PG, PG13, R, Budget
$\boldsymbol{X}_5$: Male Genre, Action, Adventure, Crime, Fantasy, Sci-Fi, Thriller
$\boldsymbol{X}_6$: Female Genre, Comedy, Drama, Mystery, Romance

For movie unit sales, we have

$\boldsymbol{X}_1$: Key variables, Constant, Family, Fantasy, Romance, Thriller, Weeks, Screens, SEN: T+22/+28

[10] See Appendix C for details and the formal proof.

$\boldsymbol{X}_2$: Sentiment, T+0, T+1/+7, T+8/+14, T+15/+21
$\boldsymbol{X}_3$: Sentiment, T-21/-27, T-14/-20, T-7/-13, T-4/-6, T-1/-3
$\boldsymbol{X}_4$: Volume, T+0, T+1/+7, T+8/+14, T+15/+21, T+22/+28
$\boldsymbol{X}_5$: Volume, T-21/-27, T-14/-20, T-7/-13, T-4/-6, T-1/-3
$\boldsymbol{X}_6$: Rating, PG, PG13, R, Budget
$\boldsymbol{X}_7$: Male Genre, Action, Adventure, Crime, Sci-Fi
$\boldsymbol{X}_8$: Female Genre, Animation, Comedy, Drama, Mystery

Statistical Logic: We first estimate the general unrestricted model (See Appendix E.4) that includes all variables by OLS. Then, we rank the variables according to their p-values from smallest to largest. For open box office (29 variables), we put the top 5 most significant variables in M_0 and group the remaining 24 variables into 6 groups of regressors of 4 in sequence. Similar for the movie unit sales (39 variables), we fix 7 most significant variables in M_0 and equally distribute the remaining 32 variables into 8 groups in sequence.

D.2 More Details on Using the LASSO for Variable Selection

Consider the linear regression model:

$$y_i = \boldsymbol{x}_{0i}^\top \boldsymbol{\beta}_0 + \sum_{j=1}^{p} x_{ji}\beta_j + u_i$$

for $i = 1, ..., n$, where $\boldsymbol{x}_{0i}$ is $k_0 \times 1$ and x_{ji} is scalar for $j \geq 1$. Let

$$\begin{aligned} \boldsymbol{\beta} &= \left[\boldsymbol{\beta}_0^\top, \beta_1, ..., \beta_p\right]^\top \\ \boldsymbol{x}_i &= \left[\boldsymbol{x}_0^\top, x_{1i}, ..., x_{pi}\right]^\top \end{aligned}$$

and define the matrices $\boldsymbol{X}$ and $\boldsymbol{y}$ by stacking observations. The OLS estimate of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}$. Consider a constrained least-squares estimate $\tilde{\boldsymbol{\beta}}$ subject to the constraint $\beta_1 = \beta_2 = ... = 0$. The LASSO estimator shrinks $\hat{\boldsymbol{\beta}}$ towards $\tilde{\boldsymbol{\beta}}$ by solving

$$\hat{\boldsymbol{\beta}}^L = \arg\min_{\boldsymbol{\beta}} \frac{1}{2n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{A16}$$

where λ is the tuning parameter that controls the penalty term. In practice, researchers either assign λ to take on a specific value or use k-fold cross-validation to determine the optimal λ. A common choice is to pick λ to minimize 5-fold cross-validation. In general, the benefits from applying the LASSO in place of OLS exist in settings where either the number of regressors exceeds the number of observations since it involves shrinkage, or in settings where the number of parameters is not small relative to the sample size and some form of regularization is necessary.

The drawback of k-fold cross-validation is its lack of computational efficiency. For exam-

ple, using five-fold cross-validation, the LASSO computation procedure is time consuming since it needs to be carried out over 200 times. The inefficiency becomes especially significant when either the sample size is big and the number of variables is large. Thus, we follow Belloni and Chernozhukov (2013) and ex-ante pick the number of explanatory variables whose coefficients will not be shrunk to zero.[11]

D.3 More Details on the Estimation Strategies

In Table 2 of the main text, we explored the forecast accuracy of the 11 estimation strategies. Details on the PMA estimator can be found in the main text. We have derived the group MMA and the group PMA strategies in Section C. Also, Section D.2 discussed the LASSO and post-LASSO strategies. The PMA post-LASSO strategy is simply the PMA model averaging process on the variables selected by the LASSO. We now provide further details the on the remaining strategies.

The general unrestricted model (GUM) strategy can be viewed as tossing everything in including the kitchen sink. By design it will have a high R-squared but the inclusion of irrelevant variables will cause losses in efficiency. The model without tweet variables (MTV) strategy in Table 2 is simply the GUM without any social media variables.

The general-to-specific (GETS) strategy modifies the general unrestricted model by removing irrelevant variables according to pre-determined criteria. We first estimate the GUM. Then, regressors with the absolute value of the t-statistics smaller than $c_\alpha = 2$ are eliminated. If multiple t-statistics are smaller than c_α, we eliminate the smallest one. The remaining regressors are retained and form a new model for the next-round test until no regressors can be eliminated. This method is also called the step-down procedure.

For the Akaike information criterion (AIC) strategy, we consider restricted variants where the AIC for a model m is defined as

$$\text{AIC}^{(m)} = 2k^{(m)} - 2\log\, L^{(m)},$$

where $\log\, L^{(m)}$ is the estimated log-likelihood and $k^{(m)}$ is the total number of regressors in model m. The model that achieves the lowest value among all of the estimated $\text{AIC}^{(m)}$ is selected by AIC.

The Mallows' model averaging (MMA) strategy is a model averaging process. The MMA criterion can be written as

$$\text{MMA}_n(\boldsymbol{w}) = \big(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{w})\big)^\top \big(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{w})\big) + 2\sigma^2 k(\boldsymbol{w}).$$

[11] Note that in Table A5 presented in subsection E.1.2, we conduct a robustness exercise that replicates the exercise presented in Table 2 of the main text which compares OLS post LASSO to model averaging post LASSO to PMA. As before for PMA model presented in the last column, the models are selected by GETS and we now allow the LASSO to select between 5 and 15 explanatory variables for both OLS and PMA with variables selected by the LASSO.

The empirical weights $\hat{\boldsymbol{w}}$ can be selected by minimizing the above criterion subject to $\boldsymbol{w} \in \boldsymbol{H_M}$. Note that the penalty term includes an unknown σ^2 that must be replaced by a sample estimate (usually provided by the largest model).

The jackknife model averaging (JMA) strategy is also known as leave-one-out cross-validation model averaging. As its name indicates, JMA requires the use of the jackknife residuals for the average estimator. The jackknife residual vector for model m can be conveniently written as $\hat{\boldsymbol{u}}_{\boldsymbol{J}}^{(m)} = \boldsymbol{D}^{(m)}\hat{\boldsymbol{u}}^{(m)}$, where $\hat{\boldsymbol{u}}^{(m)}$ is the least squares residual vector and $\boldsymbol{D}^{(m)}$ is the $n \times n$ diagonal matrix with the i^{th} diagonal element equal to $(1 - h_i^{(m)})^{-1}$. The term $h_i^{(m)}$ is the i^{th} diagonal element of the projection matrix $\boldsymbol{P}^{(m)}$. Define an $n \times M$ matrix that collects all the jackknife residuals, in which $\hat{\boldsymbol{U}}_{\boldsymbol{J}} = \left[\hat{\boldsymbol{u}}_{\boldsymbol{J}}^{(1)}, ..., \hat{\boldsymbol{u}}_{\boldsymbol{J}}^{(M)}\right]$. The least squares cross-validation criterion for JMA is simply

$$\mathrm{CV}_n(\boldsymbol{w}) = \frac{1}{n}\boldsymbol{w}^\top \hat{\boldsymbol{U}}_{\boldsymbol{J}}^\top \hat{\boldsymbol{U}}_{\boldsymbol{J}}\boldsymbol{w} \quad \text{with} \quad \hat{\boldsymbol{w}} = \underset{\boldsymbol{w} \in \boldsymbol{H_M}}{\operatorname{argmin}} \mathrm{CV}_n(\boldsymbol{w}).$$

E Additional Empirical Results

In this section, results from additional analyses that investigate the robustness of our results are presented.

E.1 Additional Evidence on the Importance of Social Media Data

We first present a subset of results from a reanalysis of the data where we only use one set of social media measures or their interactions. The first part of this analysis demonstrates that the improved performance of model averaging estimators also appears when we use a subset of the social media data utilized in the main text. Second, this analysis also provides us with intuition about what the set of sentiment measures and volume measures are respectively capturing since we complement the forecasting exercise in the main text with simple regressions that explore the amount of variation these respective sets explain.

E.1.1 OLS Results With One Set of Social Media Measures

OLS results corresponding to the GUM model for each outcome where we only control for a single measure of social media are presented in Tables A1 and A2.[12] Notice that in comparison to Table A11 where we observe that the R^2s of the estimating equations for open box and movie unit are 0.8157 and 0.8129 when both social media measures are included. Table A1 shows that adding only sentiment appears to lead to R^2s to 0.4269 and 0.5614 respectively, whereas the inclusion of volume of tweet measures leads to a larger increase in explanatory power as demonstrated in Table A2.[13]

[12] OLS estimates of the GUM model with the two social media measures are presented and discussed in subsection E.4.

[13] See Equation (A17) for details on how to calculate the R^2 for a PMA estimate.

Table A1: Models with Sentiment Only

Variable	Open Box		Movie Unit	
	Coefficient	**Std.Dev**	**Coefficient**	**Std.Dev**
Genre				
Action	-11.3892*	4.4862	-0.7671*	0.2107
Adventure	11.7951*	4.8441	0.4645	0.2585
Animation	-20.3252*	7.6094	-0.8398*	0.3702
Comedy	-0.5663	5.2920	-0.2894	0.2594
Crime	6.6253	4.2120	0.3027	0.2045
Drama	-7.2675	4.9288	-0.3723	0.2418
Family	12.4625	8.2733	0.5847	0.4022
Fantasy	9.3918	6.8777	1.0851*	0.3631
Mystery	1.1286	5.8030	-0.0703	0.2993
Romance	1.5579	5.1131	0.0987	0.2804
Sci-Fi	9.9416	5.9688	0.3667	0.2796
Thriller	7.1049	4.9221	0.1334	0.2459
Rating				
PG	10.3234	16.2355	-0.2766	2.5710
PG13	19.9404	17.7779	0.3711	2.6666
R	19.7562	17.8686	0.0805	2.6809
Core Parameters				
Budget	0.1477	0.0878	0.0091*	0.0041
Weeks	1.3065*	0.3481	0.0827*	0.0169
Screens	0.0150*	0.0035	0.0005*	0.0002
Sentiment				
T-21/-27	-0.6193	1.0721	0.0239	0.0440
T-14/-20	0.1015	1.6598	-0.0565	0.0727
T-7/-13	0.1685	2.6746	0.3593	0.1976
T-4/-6	-1.5167	3.5326	0.1466	0.1735
T-1/-3	3.2659	3.6736	-0.1102	0.2076
T+0			-0.0112	0.2075
T+1/+7			0.1275	0.2562
T+8/+14			-0.0554	0.2687
T+15/+21			-0.0058	0.3052
T+22/+28			-0.3451	0.2051
R-square	**0.4269**		**0.5614**	

* indicates the associated variable is significant at 5% level.

Table A2: Models with Volume Only

Variable	Open Box		Movie Unit	
	Coefficient	**Std.Dev**	**Coefficient**	**Std.Dev**
Genre				
Action	-0.8607	2.7540	-0.2331	0.1611
Adventure	6.8814*	2.9983	0.2922	0.1796
Animation	-10.5877*	4.5519	-0.4814	0.2659
Comedy	2.9433	3.0978	-0.0822	0.1769
Crime	3.0686	2.4889	0.0873	0.1477
Drama	-1.2196	2.8865	-0.0910	0.1635
Family	14.7771*	4.8933	0.7530*	0.2860
Fantasy	9.0517*	3.9869	1.0904*	0.2494
Mystery	3.4510	3.5182	-0.0222	0.2034
Romance	0.2018	2.9524	-0.1302	0.1897
Sci-Fi	-1.3104	3.4539	-0.0653	0.1922
Thriller	1.7918	2.8538	0.0270	0.1639
Rating				
PG	9.2063	9.4049	-0.4999	0.5336
PG13	8.9477	10.3256	-0.4076	0.5575
R	12.7364	10.3358	-0.4274	0.5524
Core Parameters				
Budget	0.1156*	0.0500	0.0048	0.0030
Weeks	0.5495*	0.2150	0.0425*	0.0128
Screens	0.0086*	0.0021	0.0003*	0.0001
Volume				
T-21/-27	-3.2987	21.2999	-1.6258	1.2530
T-14/-20	35.2492	23.5507	1.7605	1.4981
T-7/-13	-31.4952	32.3289	-3.5830	1.8292
T-4/-6	-19.9225	22.7535	0.4932	1.2895
T-1/-3	24.9256*	3.7846	0.8347	0.9864
T+0			-0.0557	0.2283
T+1/+7			0.0256	0.3320
T+8/+14			1.4796	1.0225
T+15/+21			0.2424	0.9215
T+22/+28			0.2527	0.5499
R-square	**0.8075**		**0.7867**	

* indicates the associated variable is significant at 5% level.

E.1.2 Prediction Comparison Using One Set of Measures

Throughout our analysis we find that incorporating volume and sentiment measures collected from social media improve forecast accuracy and explain a large fraction of the variation in the two Hollywood revenue outcomes. Table A3 simply carries out the simulation experiments to evaluate which approach has the greatest forecast accuracy where we only include a single type of social media data. As the results highlight, for both open box office and movie unit sales when we evaluate accuracy using the MSFE criteria, PMA has dominant performance in most of the experiments considered. The result for other thresholds are also robust.

Table A3: Results for Relative Prediction Efficiency by MSFE

Opening Weekend Box (Sentiment Only)							
n_E	GUM	MTV	GETS	AIC	JMA	MMA	PMA
10	1.3263	1.1161	1.2606	1.1390	1.0244	1.0090	**1.0000**
20	1.5184	1.2860	1.3838	1.1415	1.0133	1.0076	**1.0000**
30	1.5414	1.2153	1.3187	1.1405	1.0125	1.0132	**1.0000**
40	1.6424	1.2566	1.4099	1.1386	1.0237	1.0131	**1.0000**
Retail Video Unit Sales (Volume Only)							
n_E	GUM	MTV	GETS	AIC	JMA	MMA	PMA
10	1.3467	1.8812	1.6625	1.1505	1.0097	1.0014	**1.0000**
20	1.8179	1.7155	1.6260	1.1541	1.0084	1.0008	**1.0000**
30	2.0025	1.6407	2.6568	1.1540	1.0163	**0.9949**	1.0000
40	2.5355	1.4942	4.6391	1.1652	1.0209	**0.9889**	1.0000

Note: Bold numbers with the best performance in that simulation experiment denoted by the row of the table. The remaining entries provide the ratio of the degree of the respective forecast error metric of the estimator using specific estimation approach denoted in the column relative to results using the PMA method presented in the last column.

In addition,we do not observe additional gains from including two by two interactions between contemporaneous social media measures, which may in part be due to their high degree of collinearity.[14] We also examine the performance using only a single set of social media measures relative to the two utilized above. For open box office and movie unit sales, comparisons of the relative forecasting performance between models with only one set of social media measures (either sentiment or volume) and models with both sets is presented in Tables A4. Irrespective of the size of the evaluation set and outcome, we observe for each diagnostic used to measure forecast accuracy that (i) specification using both sets of social media variables have improved performance relative to including only one, and (ii) performance with only sentiment variables are substantially more accurate than using only volume measures.

Our analyses suggests that even with one set of social media data including the interaction between the two sets, model averaging dominates other methods. This reinforces that both opinions and the degree to which they spread are important explanatory variables for film studio revenue measures and is suggestive that film companies need to both build buzz

[14] In general, these correlations are above 0.9 and since they do not offer independent explanatory variation once we control for the main effects, it should not be a surprise that the interactions add little.

Table A4: Comparing Relative Predictive Efficiency with Different Social Media Measures

Open Box Office				Movie Unit Sales			
n_E	Sentiment	Volume	Both	n_E	Sentiment	Volume	Both
Mean Squared Forecast Error (MSFE)							
10	1.0940	2.1232	**1.0000**	10	1.0978	2.0504	**1.0000**
20	1.0955	2.0763	**1.0000**	15	1.2387	2.0013	**1.0000**
30	1.1060	2.1391	**1.0000**	20	1.3353	1.9911	**1.0000**
40	1.1191	2.3644	**1.0000**	25	1.2781	1.8675	**1.0000**
Mean Absolute Forecast Error (MAFE)							
10	1.0536	1.4652	**1.0000**	10	1.0907	1.4644	**1.0000**
20	1.0754	1.4519	**1.0000**	15	1.1007	1.4543	**1.0000**
30	1.0897	1.4874	**1.0000**	20	1.1260	1.4328	**1.0000**
40	1.0941	1.4427	**1.0000**	25	1.1538	1.4006	**1.0000**

Note: Bold numbers with the best performance in that simulation experiment denoted by the row of the table. The remaining entries provide the ratio of the degree of the respective forecast error metric of the estimator using the social media measure denoted in the column relative to results using both social media measures.

and ensure that it is positive. However, this analyses also demonstrates that while the sentiment variables play a larger role in increasing forecast accuracy, the inclusion of the volume variables does explain substantially more variation in both outcome variables. This difference is not surprising since an individual themselves is not exposed to the full volume of messages on Twitter, just the sentiment within a subset. Thus, sentiment is more likely to influence individual decisions, whereas volume can better predict aggregate outcomes.[15]

These exercises confirm both measures are needed since they capture different dimensions of product awareness and purchasing intentions. Our evidence also appears consistent with a growing behavioral and experimental economics literature that examines how opinions of others influence decision-making. For example, while there is mixed evidence on whether perceived intentions influence subsequent economic choices; much of the recent work including Bernheim, Bjorkegren, Naecker, and Rangel (2015) suggests that external information can influence decisions. Thus, it is not surprising that social media can play a role in movie purchasing decisions. Further, the heterogeneity in the effects of social media measures on the outcomes considered is also consistent with much evidence of there being substantial heterogeneity in individual decision making across settings where different individuals face identical laboratory conditions in experimental economics. As a whole, these results can be interpreted as providing support for social media marketing to influence people's intentions and subsequent purchasing decisions related to movies and is consistent with recent work in behavioral economics including Camerer, Loewenstein, and Rabin (2004), Bertrand, Karlan, Mullainathan, Shafir, and Zinman (2010) and Saez (2009) among others; that suggests that external influences can change intentions when making a suite of economic decisions.

In summary, this set of analysis demonstrates that while the sentiment variables play a larger role in increasing forecast accuracy, the inclusion of the volume variables does explain

[15]In a highly controversial study, Kramer, Guillory, and Hancock (2014) experimentally manipulated the emotional sentiment in a large group of randomly selected Facebook users' news stream. They provide evidence of emotional contagion which given the well-documented link between mood and subsequent purchasing decisions, is evidence suggestive of the pathway between our sentiment measures and outcomes considered.

Table A5: Further Comparison of the Relative Prediction Efficiency by MSFE for Open Box Office

OLS with Variables Selected by LASSO									
n_E	14	13	11	9	8	7	6	5	PMA
10	1.0990	1.1086	1.1213	1.0855	1.0137	1.0000	1.0651	1.4958	**1.0000**
20	1.1263	1.1196	1.1387	1.0574	1.1089	1.0848	1.1358	1.6853	**1.0000**
30	1.0786	1.0847	1.0911	1.0712	1.0637	1.0607	1.0994	1.4619	**1.0000**
40	1.0519	1.0867	1.0621	1.0300	1.0297	1.0781	1.0620	1.3474	**1.0000**
PMA with Variables Selected by LASSO									
n_E	14	13	11	9	8	7	6	5	PMA
10	1.0790	1.0866	1.0881	1.0854	1.0084	1.0000	1.0577	1.3685	**1.0000**
20	1.0919	1.0812	1.0929	1.0516	1.0891	1.0841	1.1221	1.6669	**1.0000**
30	1.0650	1.0665	1.0687	1.0602	1.0566	1.0587	1.0802	1.4546	**1.0000**
40	1.0044	1.0084	1.0043	1.0232	1.0267	1.0752	1.0502	1.3359	**1.0000**

Note: PMA in the last column stands for PMA method where variables are selected by GETS. Bold numbers with the best performance in that simulation experiment denoted by the row of the table. The remaining entries provide the ratio of the degree of the respective forecast error metric of the estimator using specific estimation approach denoted in the column relative to results using the PMA method presented in the last column.

substantially more variation in both outcome variables. This difference is not surprising since an individual themselves is not exposed to the full volume of messages on Twitter, just the sentiment within a subset. Thus, sentiment is more likely to influence individual decisions, whereas volume can better predict aggregate outcomes.[16]

As a final robustness check, we examine whether the conclusion of PMA post LASSO are robust to using the LASSO to select a different number of explanatory variables; hence restricting the potential models available. The results of this exercise for box office opening are presented in Table A5. Just as in Table 2 in the main text, we continue to find that PMA based on GETS has dominant performance over all of the other strategies. By discarding variables post LASSO, the total number of potential models in the second step are greatly reduced by the power of 2. However, these results continue to suggest that certain critically important potential models are dropped, which in turn induces higher MSFEs. The results of Table A5 also continue to reinforce that model averaging post LASSO yields significantly smaller MSFE than OLS post LASSO irrespective of the number of variables selected. Finally, we do not consider this expansion of an exercise for retail movie unit sales, since Table 2 clearly demonstrated the benefits of using the LASSO in the first step to undertake variable selection.

E.2 Checking Robustness to Including Seasonal and Sequel Effects

Einav (2007) discovered strong seasonality in the U.S. movie industry which appears driven by blockbuster films that are both often released during holiday weekends and have large

[16]In a highly controversial study, Kramer, Guillory, and Hancock (2014) experimentally manipulated the emotional sentiment in a large group of randomly selected Facebook users' news stream. They provide evidence of emotional contagion which given the well-documented link between mood and subsequent purchasing decisions, is evidence suggestive of the pathway between our sentiment measures and outcomes considered.

budgets. Since the movies in our sample are screened according to their budgets: 20 to 100 million, we did not expect seasonality to play as an important role in our analyses. However, since this is an empirical question, we subsequently investigated the robustness of our results to this variable as well as one that measures whether a film is a sequel/prequel in this Appendix. As detailed below, our empirical results did confirm our expectation and neither seasonality or sequel/prequel play a significant role and as such were omitted from the main analysis presented in the text.

To conduct this robustness exercise, we first cross referenced each movie release date with U.S. holidays from 2010 to 2013. We construct a new dummy variable "Seasonality", which equals 1 if the movie is released on holiday, 0 otherwise. We add the new variable to OLS regression in Appendix E.4. Empirical results are presented in Table A6. The variable "Seasonality" is highly insignificant for both open box office and movie unit sales. Also, adding Seasonality does not improve the overall R^2 for both cases.[17]

As mentioned, we also examined the effect of sequels. Similar to Seasonality, we construct a dummy variable Sequel, which equals to 1 if the movie is a sequel/prequel of a previous film. Empirical results are presented in Table A6. Like Seasonality, we find that Sequel is highly insignificant for both open box office and movie unit sales.

Table A6: Seasonality Examination

Variable	**Open Box**		**Movie Unit**	
	Coefficient	**Std.Dev**	**Coefficient**	**Std.Dev**
Seasonality	1.3532	1.9747	0.0650	0.1198
R-square	**0.8142**		**0.8096**	
Sequel	0.5412	3.8421	0.0760	0.2129
R-square	**0.8128**		**0.8088**	

E.3 Exploring Fit and the Underlying Top 5 Models

While the results in the main text show the practical advantages of using model averaging for forecasts within this industry, there are clear computational costs relative to conventional approaches. Put simply, implementing the model averaging method can be time consuming when the total number of potential models is very large. This is mainly due to the optimization routine irrespective of the software employed. To illustrate, consider the box office opening weekend example. In our data, there are a total number of 29 parameters in the general unrestricted model. Even if we fix 5 parameters in every model, it still implies a total of $2^{24} = 16,777,216$ potential models, since each model utilizes different combinations of explanatory variables and estimates the corresponding parameters.

To reduce the computational costs of the PMA estimator we considered both model screening (see Section E.5 for evidence from a robustness exercise that contrasts two procedures) and using the LASSO for variable selection in PMA post LASSO. After conducting

[17]We also expanded the definition to include if a film was released during the entire Thanksgiving to Christmas period as well as post Memorial day through July. These dates correspond to seasons with higher theatre occupancy and often coincide with releases of major films.

model screening in the main text, for open box office, we include 95 potential models in the model averaging process. For movie unit sales, this number is 56. Among these models, it may be important to understand their relative importance in the PMA estimator. Thus, we next present a summary of model averaging weights by PMA for open box office and movie unit sales in Tables A9 and A10 respectively.[18] We see in the top row that by adding across all the models the total weights of the top 5 models for each scenario account for more than 95% of the total weights. Therefore, when exploring all the weights of the potential models used for the PMA estimator, it is not a surprise in Table A7 that many the weights of many of the models is quite close to zero, even at very high quantiles of the models.

Table A7: A Summary of Model Averaging Weights by PMA

Scenario	**Mean**	**Std.Dev.**	**Quantile**				
			1%	10%	50%	90%	99%
Open Box	$M_1 = 95$						
	0.0105	0.0659	3.6E−15	6.6E−13	3.4E−9	7.2E−5	0.4359
Sum of Top 5:	**99.98%** out of total weight						
Unit Sales:	$M_2 = 56$						
	0.0179	0.0739	1.6E−20	1.2E−17	5.7E−14	0.0282	0.4928
Sum of Top 5:	**95.68%** out of total weight						

We next contrast the relative prediction efficiency of each of the top 5 models to the PMA estimator in Table A8. That is, we compare the forecasting efficiency of the top 5 models to the PMA model for each scenario using the experiment conducted in the main text. Not surprisingly, given Theorem 1 PMA shows better prediction efficiency than any of the separate models that account for substantial weight of the PMA estimator. In general, we observe models with higher weights deliver better performance than models with lower weights. In most of the experiments conducted, the top five models have better performance when the exercise set (n_E) is smaller.

Second, and returning to issues related to both fit and the importance of social media, as discussed in the main text when exploring the variables selected by the LASSO continues to demonstrate the relative importance of measures from social media in forecasting. When the LASSO respectively selects 10, 12 and 15 variables for open box office 4, 4, and 5 of which are social media measures; whereas 5, 7, and 9 are social media measures for retail movie unit sales. This indicates that among the ten variables with the strongest links to the industry outcomes considered 40 or 50% of them are obtained from social media, rather than traditional data sources that describe the characteristics of the film itself.

To further understand the importance of social media measures with this data, we also calculate the R^2 of undertaking the empirical strategy with and without sentiment variables. The R^2 for PMA is estimated following

$$R^2 = \frac{(\hat{\boldsymbol{\mu}} - \bar{y})^\top (\hat{\boldsymbol{\mu}} - \bar{y})}{(\boldsymbol{y} - \bar{y})^\top (\boldsymbol{y} - \bar{y})} = \frac{(\hat{\boldsymbol{\mu}}(\hat{\boldsymbol{w}}) - \bar{y})^\top (\hat{\boldsymbol{\mu}}(\hat{\boldsymbol{w}}) - \bar{y})}{(\boldsymbol{y} - \bar{y})^\top (\boldsymbol{y} - \bar{y})}, \tag{A17}$$

[18] Mean and standard deviation of the weighting vectors are reported. We also present their 1% to 99% quantiles.

Table A8: Results for Relative Prediction Efficiency

Opening Weekend Box						
n_E	Model 1	Model 2	Model 3	Model 4	Model 5	PMA
10	1.0443	1.0871	1.0934	1.1020	1.1561	**1.0000**
20	1.0362	1.0626	1.0520	1.0857	1.1020	**1.0000**
30	1.0253	1.0351	1.0522	1.0407	1.0648	**1.0000**
40	1.0234	1.0332	1.0435	1.0406	1.0466	**1.0000**
Retail Video Unit Sales						
n_E	Model 1	Model 2	Model 3	Model 4	Model 5	PMA
10	1.1072	1.1023	1.1126	1.1126	1.1206	**1.0000**
15	1.1028	1.1446	1.1309	1.1139	1.1394	**1.0000**
20	1.1017	1.1729	1.1606	1.1953	1.1641	**1.0000**
25	1.1152	1.1118	1.1765	1.1612	1.1281	**1.0000**

Note: Bold numbers with the best performance in that simulation experiment denoted by the row of the table. The remaining entries provide the ratio of the degree of the respective forecast error metric of the estimator using specific estimation approach denoted in the column relative to results using the PMA method presented in the last column.

where $\hat{\boldsymbol{\mu}}$ is the fitted value in general, $\bar{y}$ is mean of $\boldsymbol{y}$, and $\hat{\boldsymbol{\mu}}(\hat{\boldsymbol{w}})$ is the fitted value by model averaging. As equation (A17) indicates, the model averaging R^2 is not simply a weighted average of all R^2s by all approximation models.

E.3.1 Summarizing the Evidence

By averaging across all potential models, PMA captures the heterogeneity of all potential parameters. Therefore, it should not be much of a surprise that the fit of the PMA model in the last column of Tables A9 and A10 dominates the fit of any single model presented in the remaining columns. The R^2 statistic with and without sentiment data for the PMA model and the five models with the highest estimated weights are presented in the bottom two rows of the Tables A9 and A10. For both open box office and movie unit sales, PMA R^2s are significantly higher than any of the top five models contributing to the PMA results. This reinforces that the gains in forecast accuracy are substantial since the inclusion of social media data generally adds between 20-45% more of the explained variation. In addition, the R^2 is shown to decrease greatly if we omit Twitter variables (sentiment and volume) from any of the top five individual models, with larger declines for box office openings then retail movie unit sales. This indicates that Twitter variables have very strong explanatory power from estimating open box office, since the inclusion of these explanatory variable doubles the amount of explained variation for several of the individual models presented in Tables A9.

As a whole, our evidence presented in the main text and the online appendix suggests that social media data and model uncertainty should equally share the billing on the top of the marquee for Hollywood forecasts. This result appears consistent with a growing behavioral and experimental economics literature that examines how opinions of others influence decision-making. For example, while there is mixed evidence on whether perceived intentions influence subsequent economic choices; much of the recent work including Bernheim, Bjorkegren, Naecker, and Rangel (2015) suggests that external information can influence decisions. Thus, it is not surprising that social media can play a role in movie purchasing

Table A9: Describing the 5 Highest Weight Models: Open Box Office

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**PMA**
Weight in PMA	0.5768	0.2635	0.1210	0.0385	0.0000	
Genre						
Action						x
Adventure	x	x	x	x	x	x
Animation	x	x	x	x	x	x
Comedy						x
Crime		x		x	x	x
Drama						x
Family	x	x	x	x	x	x
Fantasy		x	x	x	x	x
Mystery						x
Romance						x
Sci-Fi						x
Thriller						x
Rating						
PG						x
PG13			x			x
R	x					x
Core						
Budget	x	x	x	x	x	x
Weeks	x	x	x	x	x	x
Screens	x	x	x	x	x	x
Sentiment						
T-21/-27					x	x
T-14/-20	x					x
T-7/-13				x		x
T-4/-6						x
T-1/-3				x		x
Volume						
T-21/-27						x
T-14/-20	x					x
T-7/-13		x	x	x	x	x
T-4/-6	x					x
T-1/-3	x	x	x	x	x	x
$\boldsymbol{R^2}$ **w/ SV.**	0.8236	0.8142	0.8172	0.8122	0.8054	0.8368
$\boldsymbol{R^2}$ **w/o SV.**	0.4027	0.4058	0.3996	0.2385	0.2385	0.4250

Note: x denotes that explanatory variable is included in the particular model, SV denotes social media data and PMA refers to our predictive model averaging method.

Table A10: Describing the 5 Highest Weight Models: Retail Movie Unit Sales

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**PMA**
Weight in PMA	0.4596	0.2709	0.1365	0.0617	0.0213	
Genre						
Action				x	x	x
Adventure	x		x	x	x	x
Animation	x			x	x	x
Comedy		x	x	x		x
Crime						x
Drama		x	x	x		x
Family	x	x	x	x	x	x
Fantasy	x	x	x	x	x	x
Mystery			x	x		x
Romance	x	x	x	x	x	x
Sci-Fi						x
Thriller	x	x	x	x	x	x
Rating						
PG						x
PG13						x
R						x
Core						
Budget	x	x	x	x	x	x
Weeks	x	x	x	x	x	x
Screens	x	x	x	x	x	x
Sentiment						
T-21/-27	x	x	x	x	x	x
T-14/-20	x	x	x	x	x	x
T-7/-13	x				x	x
T-4/-6	x	x	x	x	x	x
T-1/-3	x	x		x	x	x
T+0		x				x
T+1/+7	x	x	x	x	x	x
T+8/+14						x
T+15/+21						x
T+22/+28	x	x	x	x	x	x
Volume						
T-21/-27	x	x	x	x	x	x
T-14/-20	x	x	x	x	x	x
T-7/-13	x	x	x	x	x	x
T-4/-6						x
T-1/-3	x	x	x	x		x
T+0	x					x
T+1/+7					x	x
T+8/+14	x	x	x	x	x	x
T+15/+21					x	x
T+22/+28	x		x		x	x
$\boldsymbol{R^2}$ **w/ SV.**	0.8456	0.8337	0.8360	0.8385	0.8384	0.8672
$\boldsymbol{R^2}$ **w/o SV.**	0.4995	0.4689	0.4635	0.5465	0.5321	0.5871

Note: x denotes that explanatory variable is included in the particular model, SV denotes social media data and PMA refers to our predictive model averaging method.

decisions and likely reflects purchasing intention. As a whole, these results can be interpreted as providing support for social media marketing to influence people's intentions and subsequent purchasing decisions related to movies and is consistent with recent work in behavioral economics including Camerer, Loewenstein, and Rabin (2004), Bertrand, Karlan, Mullainathan, Shafir, and Zinman (2010) and Saez (2009) among others; that suggests that external influences can change intentions when making a suite of economic decisions.

The results in the main text and online appendix also suggest further research is needed related to both variable selection and model selection in these applications. Prior research including Wan, Zhang, and Zou (2010) has shown that it is always necessary and highly recommended to remove some poor models prior to model averaging in order to control the total number of potential models. Without doing so, not only would there be additional computational costs but a full permutation of all variables includes a huge amount of poorly constructed models will likely yield losses in efficiency and no further gains in forecast accuracy. Therefore, it is always necessary for analysts to consider which of the potential models are reasonable to include, and our analyses suggest that researchers should use algorithms from both the machine learning and econometrics literature. Finally, we point out that while concerns regarding model selection in empirical practice in this setting may appear small, since our analyses uncovered that only 5 of the thousands of models estimated accounted for over 95% of the resulting PMA estimator, the gains in forecast accuracy from PMA to any of these 5 models are non-trivial.

E.4 OLS Results

OLS results for open box office and movie unit sales are presented in Table A11. This displays the coefficients and standard errors when we use OLS to estimate the GUM model. The most striking finding is that none of the individual sentiment variables is both positively relative and statistically significant with either outcome. However, volume of Tweets immediately prior to openness is strongly related to box office revenue.

E.5 Results using other Model Screening Methods

In this section, we demonstrate the robustness of the results to using an alternative model screening method. Specifically, we consider the backward model screening (BW) procedure proposed by Claeskens, Croux, and Venkerckhoven (2006). The procedure begins with the null model and adds one variable at a time if that variable is selected by a specific information criterion. Therefore, if there are q potential variables, the BW method will pick $q+1$ nested models.

We conduct the same exercise as undertaken in Section 4 of the main text and compare the BW procedure with the GETS procedure utilized in the main text. We use Mallows' C_p as our information criterion (we did experiment with other information criteria and the results are robust and available from the authors upon request) and normalize the MSFEs according to GETS. The results are presented in Table A12. In both scenarios, model averaging using GETS as model screening method has dominant performance over the BW procedure.

Table A11: OLS Estimation Results

Variable	Open Box Office		Movie Unit Sales	
	Coefficient	Std.Dev	Coefficient	Std.Dev
Genre				
Action	-1.6740	2.7628	-0.2189	0.1653
Adventure	5.6271	2.9886	0.3039	0.1884
Animation	-12.0668*	4.4977	-0.4725	0.2710
Comedy	3.8771	3.1518	0.1363	0.1969
Crime	2.5302	2.4695	0.0697	0.1467
Drama	-2.1373	2.9139	0.0461	0.1850
Family	15.0426*	4.8439	0.8474*	0.2978
Fantasy	6.8763	4.0075	1.1017*	0.2514
Mystery	3.7158	3.4837	0.2305	0.2183
Romance	-0.4696	2.9386	-0.1805	0.2023
Sci-Fi	0.8755	3.5302	0.0084	0.1990
Thriller	2.2727	2.9620	0.2547	0.1790
Rating				
PG	9.1561	9.2292	-0.5467	1.7409
PG13	10.3972	10.1628	-0.4368	1.8098
R	14.9591	10.1766	-0.5183	1.8122
Core Parameters				
Budget	0.1297*	0.0501	0.0064*	0.0029
Weeks	0.5706*	0.2114	0.0363*	0.0132
Screens	0.0089*	0.0021	0.0002	0.0001
Sentiment				
T-21/-27	-0.6430	0.6658	-0.0563	0.0381
T-14/-20	0.7022	1.0758	0.0038	0.0670
T-7/-13	0.1360	1.5986	0.1399	0.1365
T-4/-6	-1.8274	2.0208	0.1280	0.1163
T-1/-3	2.9735	2.2055	-0.1789	0.1437
T+0			0.0407	0.1413
T+1/+7			0.2402	0.1974
T+8/+14			0.0274	0.1825
T+15/+21			0.0262	0.2060
T+22/+28			-0.3192*	0.1522
Volume				
T-21/-27	-15.4697	23.4171	-3.5547*	1.4808
T-14/-20	38.3586	26.5180	2.1758	2.0951
T-7/-13	-36.2495	33.7318	-3.8213	1.9685
T-4/-6	-8.3976	26.0235	0.6337	1.6297
T-1/-3	24.5108*	3.9924	1.8758	1.0818
T+0			-0.3035	0.2639
T+1/+7			-0.1657	0.3719
T+8/+14			2.3104*	1.1012
T+15/+21			0.3234	0.9165
T+22/+28			0.3460	0.5768
R-square	**0.8157**		**0.8129**	

* indicates the associated variable is significant at 5% level.

Table A12: Contrasting Performance between Model Screening Methods

Open Box			Unit Sales		
n_E	BW	GETS	n_E	BW	GETS
10	1.0558	**1.0000**	10	1.0274	**1.0000**
20	1.0577	**1.0000**	15	1.0434	**1.0000**
30	1.0641	**1.0000**	20	1.0382	**1.0000**
40	1.0716	**1.0000**	25	1.0432	**1.0000**

Note: Bold numbers with the best performance in that simulation experiment denoted by the row of the table. The remaining entries provide the ratio of the degree of the respective forecast error metric of the estimator using backward model screening approach denoted in the column relative to results using the GETS method presented in the last column.

References

Andreoni, J., and B. D. Bernheim (2009): "Social Image and the 50-50 Norm: A Theoretical and Experimental Analysis of Audience Effects," *Econometrica*, 77(5), 1607–1636.

Antenucci, D., M. Cafarella, M. Levenstein, C. Ré, and M. D. Shapiro (2014): "Using Social Media to Measure Labor Market Flows," Working Paper 20010, National Bureau of Economic Research.

Belloni, A., and V. Chernozhukov (2013): "Least Squares After Model Selection in High-dimensional Sparse Models," *Bernoulli*, 19(2), 521–547.

Bernheim, B. D., D. Bjorkegren, J. Naecker, and A. Rangel (2015): "Non-Choice Evaluations Predict Behavioral Responses to Changes in Economic Conditions," NBER Working Papers 19269, National Bureau of Economic Research, Inc.

Bertrand, M., D. Karlan, S. Mullainathan, E. Shafir, and J. Zinman (2010): "What's Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment," *The Quarterly Journal of Economics*, 125(1), 263–306.

Bollen, J., H. Mao, and X. Zheng (2011): "Twitter Mood Predicts the Stock Market," *Journal of Computational Science*, 2(1), 1–8.

Camerer, C. F., G. Loewenstein, and M. Rabin (2004): *Advances in Behavioral Economics,* chap. 1, pp. 3–51. Princeton University Press.

Chakravarty, A., Y. Liu, and T. Mazumdar (2009): "The Differential Effects of Online Word-of-Mouth and Critics Reviews on Pre-Release Movie Evaluation," *Journal of Interactive Marketing*, 24(3), 185–197.

Claeskens, G., C. Croux, and J. Venkerckhoven (2006): "Variable Selection for Logit Regression Using a Prediction-Focused Information Criterion," *Biometrics*, 62, 972–979.

Crawford, V. P. (2003): "Lying for Strategic Advantage: Rational and Boundedly Rational Misrepresentation of Intentions," *American Economic Review*, 93(1), 133–149.

Einav, L. (2007): "Seasonality in the U.S. Motion Picture Industry," *The Rand Journal of Economics*, 38(1), 127–145.

Einav, L., and J. Levin (2014): "Economics in the Age of Big Data," *Science*, 346(6210).

Embrey, M., G. R. Fréchette, and S. F. Lehrer (2015): "Bargaining and Reputation: Experimental Evidence on Bargaining in the Presence of Irrational Types," *The Review of Economic Studies*, 82(2), 608 – 631.

Hannak, A., E. Anderson, L. F. Barrett, S. Lehmann, A. Mislove, and M. Riedewald (2012): "Tweetin in the Rain: Exploring Societal-scale Effects of Weather on Mood," *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, pp. 479–482.

Hansen, B. (2014): "Model Averaging, Asymptotic Risk, and Regressor Groups," *Quantitative Economics*, 5, 495–530.

Hansen, B. E. (2007): "Least Squares Model Averaging," *Econometrica*, 75(4), 1175–1189.

Holbrook, M. B. (1999): "Popular Appeal Versus Expert Judgments of Motion Pictures," *Journal of Consumer Research*, 26(2), 144–155.

Hsu, C. L., and J. C. C. Lin (2008): "Acceptance of Blog Usage: The Roles of Technology Acceptance, Social Influence and Knowledge Sharing Motivation," *Information & Management*, 45(1), 65–74.

Karabulut, Y. (2013): "Can Facebook Predict Stock Market Activity?," *Working Paper*.

Kim, J., and D. Pollard (1990): "Cube Root Asymptotics," *The Annals of Statistics*, 18(1), 191–219.

Kleef, G. A. V., C. K. D. Dreu, and A. S. Manstead (2010): "An Interpersonal Approach to Emotion in Social Decision Making: The Emotions as Social Information Model," *Advances in Experimental Social Psychology*, 42, 4596.

Kramer, A. D. I., J. E. Guillory, and J. T. Hancock (2014): "Experimental Evidence of Massive-scale Emotional Contagion through Social Networks," *Proceedings of the National Academy of Sciences*, 111(24), 8788–8790.

Lee, D. (2008): "Game Theory and Neural Basis of Social Decision Making," *Nature Neuroscience*, 11, 404 – 409.

Mishne, G., and N. Glance (2006): "Predicting Movie Sales from Blogger Sentiment," *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*.

Mislove, A., S. L. Jørgensen, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist (2011): "Understanding the Demographics of Twitter Users," *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pp. 554–557.

Saez, E. (2009): "Details Matter: The Impact of Presentation and Information on the Take-Up of Financial Incentives for Retirement Saving," *American Economic Journal: Economic Policy*, 1(1), 204–28.

Toole, J. L., Y.-R. Lin, E. Muehlegger, D. Shoag, M. C. González, and D. Lazer (2015): "Tracking Employment Shocks Using Mobile Phone Data," *Journal of The Royal Society Interface*, 12(107).

Wan, A. T., X. Zhang, and G. Zou (2010): "Least Squares Model Averaging by Mallows Criterion," *Journal of Econometrics*, 156(2), 277 – 283.