Identification of Trends in Consumer Behavior through Social Media

David Alfred Ostrowski
System Analytics
Research and Innovation Center
Ford Motor Company
dostrows@ford.com

Abstract

Social Media has frequently been leveraged for the purpose of anticipating trends. This paper presents a methodology to identify trends of consumer behavior through Semantic Filtering. The process begins by applying keyword-based filtering on a ground topic, proceeding to filter on terms related to anticipation of a purchase. Next, semantic categories are considered to filter out messages that are inconsistent with the desired signal. Following, Fisher Classsification is applied to identify consumer behavior. We apply this procedure to the goal of modeling vehicle purchase behavior with data acquired from Twitter. Results demonstrate that consumer demand can be determined within the Twitter firehose, providing a source of data to contribute to forecasting efforts as well as Customer Relationship Management (CRM).

1. INTRODUCTION

Internet-based activities have demonstrated to be highly reflective of real-world situations. Correspondingly, monitoring internet-based data sources have demonstrated to support the identification of trends across many subjects. By considering these sources from the perspective of a social network, semantics have been applied for substantial use in trend identification [1]. Many of these sources including Social Media have demonstrated advantages for use in market prediction due to the fact that they are frequently updated and include very unbiased results. Due to the increased popularity of these sources and subsequent volume they have become a form of disruptive technology, displacing traditional information media [2].

Analytics around internet-based data sources have been leveraged to support the areas of forecasting, public relations and CRM [3][4][5]. This has been supported through areas including Machine Learning and Natural Language Processing. By leveraging classification related to sentiment, internet sources have allowed for the means to semi-automate the engagement process with customers thus allowing greater utility than previous forms of engagement which include email, customer groups and surveys[6].

Due to its minimalistic characteristics and ease of data collection, we have targeted the Twitter micro-blogging service as our data source. Twitter volume has increased in recent years, presenting a formidable challenge having over 500M active users, generating 430M tweets and handling over 1.6M search queries per day [6]. Through such large participation, Twitter has been able to support the identification of trends across general topics that affect very large populations. The challenge that exists is to extract signals that assist in the definition of consumer behavior within larger macroeconomic activity. Such signals can also support extraction of underlying topics that will support a greater understanding in marketing and customer relationships.

In this paper we investigate the application of these techniques to determine a more focused level of behavior that exists within larger trends. Specifically, we are interested in the identification of stages of consumer activity as it relates to an individual product line. Here, the focus is to identify specific consumer demand for an item towards application of sales forecasting as well as CRM.

In the next section, we present a survey of related work. Section three continues with the discussion of our proposed methodology. Section Four presents our test case and Section Five presents our conclusions.

2. CURRENT RESEARCH

Many techniques adapted from categorization of unstructured data have been applied to Social Media. Techniques range from rule-based and semantic evaluation to Machine-Learning based methods. Being a popular source within Social Media, Twitter has been leveraged for the determination of general trends. By comparing Twitter information with associated sentiment (Tumasjan et.al.) was able to determine that Twitter messages could provide an accurate portrayal of the political landscape [6]. Through improving on the quality of a Twitter document collection and through the incorporation of sentiment analysis (Sang et. al.) supported predictions based on entity counts matching performance of traditionally obtained opinion polls [7]. (Bermingham et.al.) demonstrated predictivity in social analytics through the use of both volume-based measurements and sentiment analysis as explored over a variety of sample sizes, time periods and quantitative methods [8].

(Stewart et. al.) examined aggregate daily Twitter keyword volumes to predict aggregate current spending. They demonstrated that weekday Twitter keyword volume, current spending, and weekday spending norms all have significant value allowing for prediction of short-term consumer spending [9]. Within the entertainment industry, (Asur et. al.) leveraged social media to predict real-world outcomes by constructing a model that forecasts box-office revenues for movies through the construction of a linear regression model, outperforming other market indicators such as Hollyworld Stock Exchange [10]. (Gruhl et. al.) showed how to automatically generate queries for mining logs in order to predict spikes in book sales [11]. (Huberman et. al) studied the social interactions on Twitter to reveal that the driving process for usage is a sparse hidden network underlying the friends and followers, while most of the links represent meaningless interactions [12].

Researchers have also been successful in the past in identifying and tracking macro-level events including seasonal illnesses through Social Media. (Signorini et. al.) used Twitter to identify and track levels of disease activity and public concern in the U.S. during the influence of the H1N1 panademic [13]. This was accomplished by filtering keywords that would contain disease names as well as public concern keywords with the results showing that Twitter can be used as a measure of public interest or concern around health related events. (Chew et. al.) collected twitter messages containing keywords related to the H1N1 virus during the 2009 H1N1 outbreak and thus validated twitter as a tracking system for public attention [14]. (SonDoan et.al.) was able determine a novel algorithm applied to the filtering of datasets by application of a semantic featureset including consideration of negation , hashtags, emoticons, humor and geography thus resulting in a improvement over earlier taxonomy based approaches leveraged to semantic filtering [15].

Work in Twitter also involves the evaluation of trends as they relate to social networks. (Asur et.al.) identified that relationships between members of a social network play a substantial role in creating trends [16]. They also determined that trends are also based on their influences with the passivity of users predicated on their information forwarding activity. Their study also showed that that the correlation between popularity and influences was weaker than might be expected.

A number of methods have leveraged machine classification in order to generate trends among unstructured and social-network based data collections. Among lexical-based approaches, (Xia et. al.) analyzed attribute sentiment co-locations to approximate reasonable generalization abilities [17]. Working with both lingual analysis as well as classification and clustering methods, Shandilya and Jain worked towards the extraction of knowledge utilizing a hybrid approach [18]. More sophisticated multi-pass efforts have been developed recently with an example being Xu and Kit who performed examinations of opinions at different levels (course/ fine) of a document [19].

(Wang et. al.) relied on Latent Dirichlet Allocation (LDA) in which a topic was associated with a continuous distribution over timestamps and for each generated document, the mixture distribution over topics was influenced by both word co-occurences and the document's timestamp. This was leveraged to interpret trends from email , research papers and state-of-the-union addresses[20]. Padman and Airoldi proposed an approach to extract sentiments from unstructured text through the means of applying a two-stage Bayesian algorithm that is able to capture the dependencies among words and at the same time find a vocabulary that is efficient for the purpose of mining algorithms [21]. Other approaches for trend identification include the application of centrality calculations within a Social Network as well as search engine technology algorithms such as PageRank algorithm such as that employed by (Corely et. al.) which was able to determine that Influenza related blogging trends have a significant correlation to the US Fall 2008 flu season [22]. They were also able to identify WSM Influenza-related communities that share flu-postings which could

broker or disseminate information in the case of a severe outbreak or Influenza epidemic.

3. METHODOLOGY

We begin with the overview of our methodology by considering an internet-based data resource (Twitter) as our starting point. We are interested in a demand signal of an individual product for the application of sales forecasting and trend prediction. To this means we consider the theoretical journey made by consumers noted as the Awareness, Interest, Desire and Action (AIDA) model, also known as the 'Purchasing Funnel'[23]. With this in mind, our attributes are to match the consumer emotion of desire, in anticipation of the purchase event in the short term (within a month). As noted in Figure 1, our methodology employs two separate levels of direct filtering (empirical and knowledge) combined with a classifier to support the extraction of a signal.

Our process starts with the consideration of single ground truth keywords that are applied to the raw data collection (Twitter Firehose). These are designated as words for unique identification of a product. Here, a tradeoff is presented between containment of the maximum amount of potential information and precision of matching the given subject.

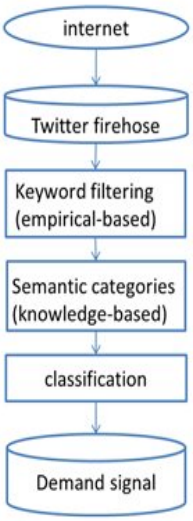

Figure 1. Overall methodology

In our second stage, we are interested in keywords that are directly related to the anticipation of a purchase event. Not having found a suitable ontology or taxonomy, a corpus was created from taking high frequency words from messages that were most likely to indicate the purchase of a product. This was accomplished by hand selecting Twitter messages indicating a high level of desire among consumers for a specific product. Employing a "try and test" method we were able to derive a subset of words that indicated a collection of Twitter messages who's volumes were able to correlate to sales. Here, we applied the following keywords indicating such terms including "test drive", "purchase", "buy" and "insurance" as applied to our subject matter data collection.

Next, we examined three specific semantic categories to apply our filtering. These included negation, hashtags and humor. Negation was examined first. In this step, we applied a natural language processing approach to break down the "demand based" indicators between positive and negative demand through the application of the Python Natural Language TookKit (NLTK). This process examined basic part-of-speech (POS) taggings in each sentence and applied a set of regular expressions to support identification of both direct and indirect relationships between words supporting negation (including 'not' and 'never') and the ground truth keywords.

Hashtags were next considered as a means of filtering content. A hashtag is considered as a token beginning with the symbol '#'. As such it is considered as a community-driven comment for the addition of context and metadata to tweets Here, a taxonomy of hashtags was developed that were both related and unrelated to the topic of demand. Any topics that were unrelated to the Twitter data were eliminated. The related topics could be used to consider the weighting of the messages. Humor was also considered at the word level with specific terms associated with common jokes or humor were eliminated.

For the classification step, a training set of documents was manually devised from samples of Twitter messages that indicated a high level of desire among users in a product. Towards support of semantics identification we employed the Fisher classification method. This approach allowed for a higher level of flexibility in identification of semantic combinations as opposed to the standard naïve Bayesian algorithm by supporting a normalized method of classification. To perform such normalization the method relies on three calculations:

clf = Pr(feature| category)

$$\text{Freq sum} = \sum_{i\ldots}^{n} \Pr(Feature \mid Category)$$

cprob = crlf / (clf + nclf)

Clf is determined as a conditional probability that a document fits into a category, given a particular feature. This is in contrast to the Bayesian approach

of considering the number of documents with features divided by the total number of documents (with that feature). By approximating the feature to category level, it takes into account of receiving far more documents in one category than another. To normalize, the probability is divided by the frequency sum. The Fisher method continues by multiplying all the probabilities together, taking the natural log and applying the inverse chi function to obtain a probability. Figure 2, provides a detailed overview of the three step filtering process.

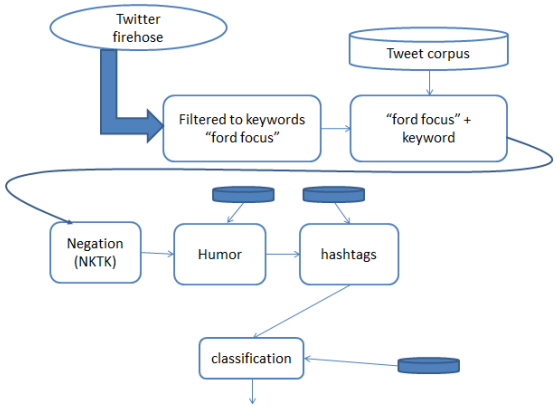

Figure 2. Filtering Level Detail

4. IMPLEMENTATION

The goal of our case study was to characterize the demand for the Ford Focus sampled from the Twitter firehose for the time period of October 1, 2011 to September 31, 2012. The data collections were considered on a monthly basis. Following our model of the purchasing funnel, the focus was to characterize the anticipation immediately before the purchase event. Our empirical filtering begins with the determination of ground truth keywords designated as any message that contains both the words "ford" and "focus" supporting over a 99% precision rate. While these words eliminated messages which included statements such as "focus" without the inclusion of "ford", precision fell to under 20% when "ford" was omitted, making it necessary to utilize the keyword combination. The volumes of such filtering are presented in Figure 3, generating a .26 correlation to vehicle sales.

The second stage of our empirical filtering consisted of the establishment of a manual taxonomy from a sampled collection of 4000 Twitter messages (filtered accordingly to the initial ground truth keywords) and associated strongly to the anticipation of the purchase event. This included the following keywords which included (buy, purchase, love, insurance). The highest performing set of these keywords is presented in Figure 4, generating a correlation to Ford Focus sales as presented in figure 6 of .744.

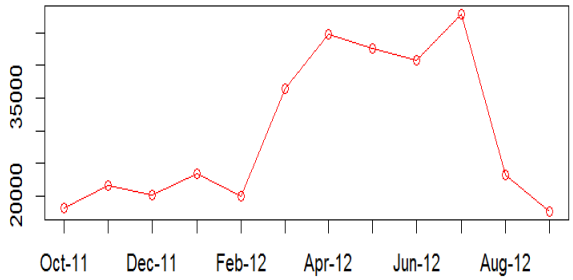

Figure 3. Ground Truth Keyword Volumes (Ford Fusion) for the period Oct-11 to Sept-12

Next, semantic categories were applied with each category supporting successive filtering. The largest proportion of these categories was humor followed by hashtags and negative. The application of the semantic categories such as humor included the matching against keywords known to be associated with messages around humor including "Like Ford I got Focus" along with references between the words "focus" and "attention" The results of which are presented in the semantic category section in figure 5. This resulted in only a marginal increase in correlation (.7575) as few messages at this point were filtered.

The classification step was trained (positively) from a manually derived collection of messages indicating demand and also trained (negatively) against messages that indicated advertising for new or used vehicles. This resulted in a minimal reduction of the overall collection increasing the correlation to .76 (figures 3 and 5).

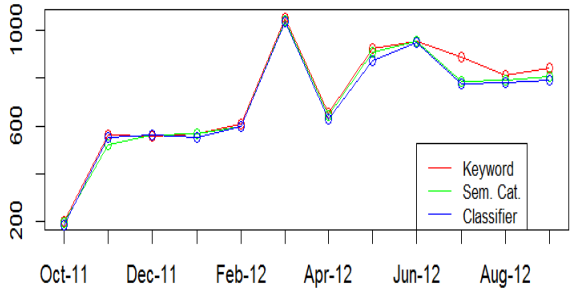

Figure 4. Generated message volumes for Keyword, Semantic Category and Classifier – enabled filtering for period Oct-11 to Sep-12.

To support our work in forecasting we compared these results to Google Trends (another source of web-based information utilized for forecasting and indication of trends) for the same period. We examined Google Trends searches for the terms "ford focus" , "ford" and "ford dealers", under the United

States and Vehicle Sales category. The correlation for our data sample fell short of the filtered data at .71,.59 and .42 respectively with a correlation of only .46, .37 and .32 to our best filtered collection (the final stage).

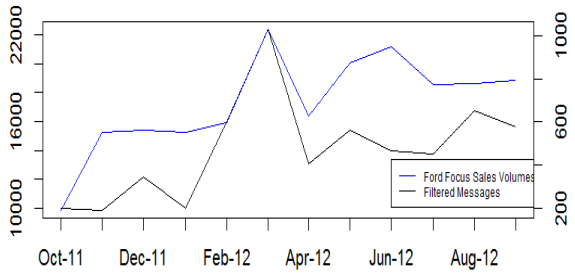

Figure 5. Comparison of Ford Fusion Sales volume to Filtered Messages Volume for period (Oct-11 to Sep-12)

category	count	correlation
"ground truth" keywords	356603	0.263535
"empirical" keywords	8614	0.7443
semantic categories	8392	0.7571
classifier	8283	0.76

Figure 6. Correlation of levels of filtering to Sales Volumes

category	correlation	Correlation to SM-f
Google Trends ("Ford Fusion")	0.71	0.46
Google Trends ("Ford Dealers")	0.59	0.37
Google Trends ("Ford")	0.42	0.32

Figure 7. Correlation of Google Trends to Sales and Semantically Filtered Data.

5. CONCLUSION

We have proposed a methodology for the monitoring of trends at an individual product level. Our process considers the application of unstructured data through filtering to determine a demand indicator to reflect consumer behavior. The results indicate that a strong demand signal can be generated for application in forecasting and CRM. Compared to another popular web-based indicator (Google Trends) our model performs well, providing higher correlation. Our filtered signal also maintains a low correlation to this attribute (.46) indicating that it could perform well as a complementary parameter. While our core strategy involves the application of Twitter this methodology could be extended across any number of social media platforms as well as collections of unstructured data.

Future work includes additional methods of comparison between keyword sets with the possible inclusion of Machine Learning technologies to assist in the determination of optimal or near-optimal combinations. The weighting of keywords in filtering as well as expanded training sets in classifiers can also assist in supporting a stronger signal that will enable the performance of a forecasting model.

6. REFERENCES

[1] Gloor, P.A.; Krauss, J.; Nann, S.; Fischbach, K.; Schoder, D.; "Web Science 2.0: Identifying Trends through Semantic Social Network Analysis", Computational Science and Engineering, 2009. CSE '09. International Conference on Volume: 4 pp: 215 – 222

[2]Kwak, Haewoon, Changhyun Lee, Hosung Park, and Sue Moon. "What is Twitter, a social network or a news media?." In Proceedings of the 19th international conference on World wide web, pp. 591-600. ACM, 2010.

[3] Ostrowski, D. A. , "Predictive Semantic Social Media Analysis, IEEE International Conference on Semantic Computing , ICSC , 2011

[4] Baird, Carolyn Heller, and Gautam Parasnis. "From social media to social customer relationship management." Strategy & leadership 39, no. 5 (2011): 30-37.

[5]Ang, Lawrence. "Community relationship management and social media."Journal of Database Marketing & Customer Strategy Management 18, no. 1 (2011): 31-38.

[6]Tumasjan, Andranik, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe. "Predicting elections with twitter: What 140 characters reveal about political sentiment." In Proceedings of the fourth international aaai conference on weblogs and social media, pp. 178-185. 2010.

[7]Sang, Erik Tjong Kim, and Johan Bos. "Predicting the 2011 dutch senate election results with twitter." EACL 2012 (2012): 53.

[8]Bermingham, Adam, and Alan F. Smeaton. "On using Twitter to monitor political sentiment and predict election results." (2011).

[9] Stewart, Justin, Strong, Homer, Parker, Jeffrey, Bedau, Mark A., Twitter keyword volume, current spending, and weekday spending norms, predict consumer spending., 2012 IEEE 12th International Conference on Data

[10] Asur, Sitaram, and Bernardo A. Huberman. "Predicting the future with social media." arXiv preprint arXiv:1003.5699 (2010).

[11] Daniel Gruhl, R. Guha, Ravi Kumar, Jasmine Novak and Andrew Tomkins, The predictive power of online chatter. SIGKDD Conference on Knowledge Discovery and Data Mining, 2005

[12]Benardo A. Huberman, Daniel M. Romero, and Fang Wu, Social networks that matter: Twitter under the microscope. First Monday, 14(1), Jan 2009

[13] A. Signorini, A. M. Segre,and P. M. Polgreen, "The use of Twitter to Track Levels of Disease Activity and

Public Concern in the U.S. during the Influenze A H1N1 Pandemic" PLoS ONE, vol 6. , no. 5, p. e19467 05 2011.
[14] chew and eysenbach "Pandemics in the age of twitter: content analysis of tweets during the 2009 H1N1 outbreak PLos One, vol5, no 11. P e14118, 11 2010
[15] Doan, Son, Ohno-Machado,Lucia, Collier, Nigel, Enhancing Twitter Analysis with Simple Semantic Filtering: Example in Tracking Influenze-Like Illnesses, IEEE 2nd Conference on Healthcare Informatics, Imaging and Systems Biology, 2012
[16]Asur, Sitaram, and Bernardo A. Huberman. "Predicting the future with social media." arXiv preprint arXiv:1003.5699 (2010).
[17] Yun-Qing Xia; Rui-Feng Xu; Kam-Fai Wong; Fang Zheng; The Unified Collocation Framework for Opinion Mining, Machine Learning and Cybernetics, Hong Kong, Volume 2, 19-22 Aug. 2007 pp 844 – 850
[18] Shandilya, S.K.; Jain, S.; Automatic Opinion Extraction from Web Documents, Computer and Automation Engineering, 2009. ICCAE '09. International Conference on 8-10 March 2009 pp. 351 – 355
[19]Ruifeng Xu; Chunyu Kit; Corse-fine opinion mining Machine Learning and Cybernetics, 2009 International Conference on Volume 6, 12-15 July 2009 pp 3469 - 3474
[20] Wang, Xuerui, and Andrew McCallum. "Topics over time: a non-Markov continuous-time model of topical trends." In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 424-433. ACM, 2006.
[21]Airoldi, Edoardo, Xue Bai, and Rema Padman. "Markov blankets and meta-heuristics search: Sentiment extraction from unstructured texts." Advances in Web Mining and Web Usage Analysis (2006): 167-187.
[22]Corley, C., Armin R. Mikler, Karan P. Singh, and Diane J. Cook. "Monitoring influenza trends through mining social media." In International Conference on Bioinformatics & Computational Biology, pp. 340-346. 2009.
[23] Barry , Thomas, 1987, The Development of the Hierarchy of Effects: An interesting Historical Perspective, Current Issues and Research in Advertisting, 251- 295