Characterizing Brand Advertising Strategies on Twitter

Shana Dacres, Hamed Haddadi, Matthew Purver
Cognitive Science Research Group
School of Electronic Engineering and Computer Science
Queen Mary University of London, UK
firstname.lastname@eecs.qmul.ac.uk

ABSTRACT

Social media have substantially altered the way brands and businesses advertise. Through the use of engagement in Online Social Networks, brands are enjoying a more versatile and dynamic channel for advertisement than they would otherwise have had using traditional media (e.g., TV and radio). Recently, Twitter has introduced two advertising mechanisms as channels of influence: *promoted tweets*, and *promoted trends*.

Using data collected from Twitter, we analyze how engagement and sentiment in promoted content spread over a 10-day period. We use Machine Learning and Natural Language Processing techniques to gather focused, relevant datasets, and to accurately gauge sentiment. We find that in user interactions, promoted tweets lead to higher positive sentiment than promoted trends. However, promoted trends do pay off in response volume. We observe the highest percentage of hashtag adoption on the first day of the campaign, with engagement levels for the brand and promoted content falling considerably thereafter. Our methodology highlights the importance of using robust methods for detecting the topic and sentiment of social texts, rather than relying on simple keyword- or frequency-based metrics.

Keywords

Online Social Networks, Advertising, User Engagement

1. INTRODUCTION

Online Social Networks (OSNs) such as Facebook, Twitter, and YouTube have emerged as highly engaging marketing and influence tools, increasingly used by advertisers to promote brand awareness and catalyze word-of-mouth marketing. Researchers have also long recognised the effectiveness of OSNs as a rich source for understanding the spread of information about the real world [16]. For example, Asur *et al.* [1] analyzed Twitter messages (*tweets*) to predict box-office ratings for newly released movies. Their findings shows that OSNs can be used to make quantitative predictions that outperform those of markets forecasts, by focusing on the *sentiment* expressed in the tweets. Brands also now recognise the potential of OSNs for gathering market intelligence and insight. In 2012, Twitter announced that 79% of people follow brands to get exclusive content.[1] This provides the opportunity for brands to participate in real-time conversations to listen to and engage users, respond to complaints and feedback, drive consumer action and broadcast content.

In this study we examine the content and volume of users' brand engagement on OSNs to determine the effect of choice of promotion channel on a brand's influence. We do this by analysing the engagement level of Twitter users, their adoption of brand hashtags, and the sentiment they express, to determine the similarities and differences between two separate advertising strategies on this network: *promoted tweets*, and *promoted trends*. We pose a number of questions regarding brands and advertising on OSNs: How does the sentiment for a promotion strategy spread over time? What are the engagement levels for each day of promotion? What is the engagement level (e.g. *retweets* and *mentions*) for promoted brands and how do these affect the sentiments expressed towards a brand?

In order to answer these questions, we use Twitter's Streaming API service to collect engaged users' profiles and tweets in regards to promoted influences (tweets and trends) over a busy 10 day shopping period for a selection of brands across different industries. We use Machine Learning (ML) techniques to accurately filter the tweets for topical relevance, a task which simple keyword-based methods could not achieve. We then use established Natural Language Processing (NLP) tools to classify the tweets by sentiment (positive, negative, or neutral). We then use this data to establish the driving factors behind the success of promoted influences and differences between advertising strategies.

2. RELATED WORK

There have been a number of recent studies about individuals' influence on Twitter [3], and the effectiveness of online advertising [4, 2], but little attention has

[1] `http://advertising.twitter.com/2012/05/twitter4brands-event-in-nyc.html`

been paid to identifying the driving factors behind a brand's influence on their social audience (although it has been noted that brand names are more important online for some categories [6]). Cheung *et al.* [5] examined the way information spreads differently within social networks as opposed to word-of-mouth (WOM) broadcasting, by focusing on electronic word-of-mouth (eWOM), showing comprehensiveness and relevance to be the key influences of information adoption. The closest work to ours in understanding brands on Twitter is the study by Jansen *et al.* [9], who found that 20% of tweets that mentioned a brand expressed a sentiment or opinion concerning that company, product or service. Here, we examine and compare the effects of the promotion strategies available to brands specifically for advertising effectiveness on Twitter (see Section 3).[2]

In an important study on the spread of hashtags within Twitter, Romero *et al.* [18] used over 3 billion tweets 2009-2010 to analyze sources of variation in how the most widely used hashtags spread within its user population. Their results suggested that the mechanism that controls the spread of hashtags related to sports or politics tends to be more persistent than average; repeated exposures to users who uses these hashtags affects the probability that a person will eventually use the hashtag more positively than average. A limitation of their paper is that they only concentrated on hashtags that succeeded in reaching a large number of users. In regards to the focus of promoted influences within Twitter, this raises the question; what distinguishes a promoted item that spreads widely with mainly positive sentiment, from one that fails to attract attention or has mainly negative sentiment posts? Our study aims to answer this by examining the sentiment and spread of tweets in relation to that brand's promoted item.

Sentiment analysis has been approached across many domains, including products, movie reviews and newspaper articles as well as social media (see e.g [14] for a comprehensive overview). Typically, the methods employed depend either on existing language resources (e.g. sentiment dictionaries or ontologies) or on machine learning from annotated datasets. The former can provide deep insight, but are somewhat inflexible in the face of the non-standard and rapidly changing language used on OSNs, for which few suitable linguistic resources currently exist. The latter are more scalable and can be trained on relevant data (e.g. [11]), but generally depend on large amounts of manual annotation (expensive and often problematic in terms of accuracy). However, some approaches leverage the existence of implicit labelling in the datasets available, to avoid the necessity for manual annotation: for example, user ratings provided with movie or product reviews [15, 4]); or author conventions such as emoticons and hashtags on OSNs [7, 13, 17]). Here we use an existing tool derived using this latter approach (see [17]) and available free online.[3]

Industry	**Promotion type**	**Brand**
Electronics	Promoted tweet	International CES
	Promoted tweet	SONY
	Promoted trend	Nintendo UK
Travel	Promoted tweet	Marriot
Entertainment	Promoted tweet	BBC One
Automobile	*Promoted trend*	Vauxhall
Heath Care	Promoted tweet	Paints like Me
Retail	*Promoted trend*	ASOS
	Promoted trend	PespiMax
	Promoted tweet	JRebel
Telecomms	*Promoted trend*	O2 Network

Table 1: Industry sectors and sample brands

3. DATA COLLECTION

We set up a crawler to use the Twitter Streaming API[4] to collect the tweets of interest and all associated metadata (e.g., ID, username, user's social graph), with details stored in a MySQL database. In this section we briefly describe our dataset and data collection strategy.

Identifying promoted brands

Twitter distinguishes promoted tweets and trends by the use of a *Promoted* tag. We collected tweets across six different industry domains, ranging from entertainment to health-care. For each promoted item, the brand name was used to crawl Twitter for tweet data posted in English for a 10 day period. If the promoted item also included a hashtag, the hashtag was also included in the parameters of the crawl's GET function. This included all tweets that contained keywords such as `@BrandName`, `#BrandName`, `BrandName`, `#PromotedHashtag` and other brand related terms. These parameter values were selected to keep the dataset both relevant to brand-related tweets, and also manageable for searching purposes. Followers and following information was also tracked on a daily basis for each brand.

Details of the selected brands and their promoted type are provided in Table 1. Given that we were interested in promoted items for branding purposes, a range of different brands from different industries were selected. The aim was to include both major, and small brands when selecting promoted items. In addition, a major brand and a small brand enable a comparison of sentiment while weakly controlling for follower count.

Dataset

We identified different industry's promoted items for 10 day periods between 17th December 2012 and 7th Jan-

[2] Using our dataset we can not possibly determine whether an advertising campaign leads to actual clicks or sales.

[3] `http://chatterbox.co/api/`

[4] `https://dev.twitter.com/docs/streaming-apis`

uary 2013. We used non-parallel crawling periods in order to avoid the query limits set by the Twitter API. In total, around 180,000 individual tweets were collected by crawling Twitter continuously, excluding December 21^{st} 2012 when there was a 6 hour outage in the crawler API. The crawler collected tweets from around 120,000 different Twitter users engaged in spreading the promoted tweets and trends. Tweets across all topics and with no geographical limits were gathered, as long as they featured the brand's name/hashtag.

In order to remove noise and bias in analysis caused by spam tweets, we removed users who had posted the exact same tweet more than 20 times, along with their tweets. Twitter users, tweets and tweet timestamps were also cross-analysed to check for spamming accounts. In one case a single user was removed for adding over 8,000 spam tweets to the database.

4. TEXT PROCESSING & CLASSIFICATION

In this section we present the details of our tweet classification (using ML) and sentiment analysis (using existing NLP tools).

4.1 Topic Classification

One of the major challenges during cleaning the dataset and removing spam was ensuring topic relevance. Our expectation was that this would not be an issue: as in much previous work, our study is looking at all sentiment expressed towards the brands, as long as the tweet matched the parameters of the tweet selection as explained in Section 3. However, whilst sampling tweets for spammers, a general problem surfaced. We found that our keyword-based approach was too limited to accurately identify tweets referring to a particular brand, *O2* (a UK mobile telecommunications provider and network). Our parameters for collecting tweets for this brand were to match tweets containing `O2WhatWouldYouDo` and `O2` (the hashtag being promoted was `#O2WhatWouldYouDo` and `@O2` is the official brand Twitter handle). Over the 10 day period, 90,000 tweets were collected that matched these keywords. However, examining a random sample of 200 tweets from this dataset showed that over 70% were not referring to the O2 Network brand; many were referring to the *"O2 Academy"* (a chain of concert venues), the *"O2 Arena"* (a dome-shaped monstrosity in London), or other senses of *'O2'* such as oxygen. We also noticed that Twitter users have recently established a new way of using the letter sequence *'O2'* as a replacement for the letters *'to'*: e.g. "`@CokeWave_Thang What Picture You Want Me O2 Put As My BackGround`", "`what im goin o2 do o2day`". Experiments with boolean combinations of `O2` with other keywords were not successful. A major challenge therefore becomes to filter out non-brand-related tweets automatically: the problem is not trivial, given the variability and unpredictability of language, vocabulary and spelling on Twitter, and the short length of tweets (up to 140 characters); and manual removal of approximately 70% of large datasets is prohibitively labour-intensive.

We therefore approached this as a text classification problem and investigated various supervised machine learning approaches using the Weka toolkit [8]. First, we performed a pilot study over a 200-tweet development set to determine a suitable feature representation and classification method; the data was manually labelled as O2-related or otherwise to give a binary decision problem. We tested a variety of classifiers including Naive Bayes, Naive Bayes Multinomial, ID3, IBK and J48 decision trees; features were based on the tweet text using a standard bag-of-words representation (see e.g. [10]) with various scaling methods,[5] with the addition of user ID and date of tweet. Given the small size of the dataset, we restricted the feature space to be based on the most common 100 words. We also tested using a simple manual keyword-based filter to remove some common negative instances (using keywords *arena, academy*, etc) before training (see "manually filtered" results in the figures). Tests were performed using ten-fold cross-validation in order to simulate performance on unseen data. Best performance (overall accuracy) was obtained using only bag-of-words text features, with stopwords removed and a TF-IDF weighting, after manual filtering. The best performing classifiers in cross-validation were J48 and Naive Bayes (NB), with 71% and 91% accuracy respectively. We then compared their performance on a held-out test set: the NB model outperformed the J48 model with 84% accuracy compared to 71% for J48, with training and prediction also noticeably faster for NB (the tree structure of the J48 model made it very slow with larger training sets).

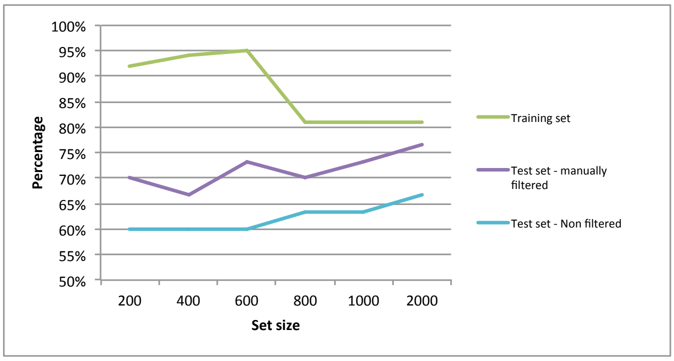
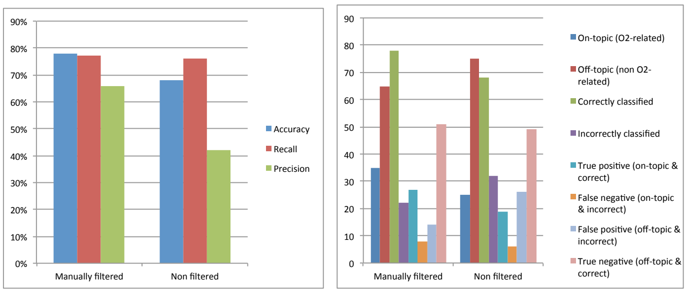


Figure 1: NB accuracy with increasing training data.

To determine a suitable training set size, we then varied the training set while testing performance on a held-out test dataset of 30 manually labelled tweets.

[5] We used Weka's `StringToWordVector` filter for text feature extraction and scaling.



(a) Overall results

(b) Detailed results per class

Figure 2: Classification results for Naive Bayes method.

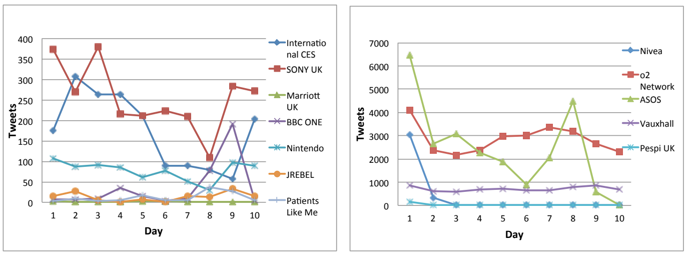
Increasing training set size improved performance (see Figure 1): we tested up to a 2,000-tweet training set; while the curve suggests performance may improve beyond this point, the accuracy on the held-out test set is approaching that on the training set so large improvements are unlikely. The NB classifier trained on 2,000 tweets was therefore used for the experiments below. Figures 2a and 2b display results when tested on a larger, unseen, randomly selected test set of 100 tweets; the version with manual filtering achieves 78% accuracy, 77% recall and 66% precision.

4.2 Sentiment Analysis

Having identified tweets with relevant content, we now required a method for sentiment analysis – determining the positive or negative stance of the writer. As discussed in Section 2 above, many methods for sentiment detection exist, with the major distinction being between lexicon-based and machine learning-based approaches. We examined existing tools for Twitter sentiment analysis using both of these approaches in order to determine the most suitable for our data.

As a lexicon-based tool we used SentiStrength [19]. This method uses a predetermined list of words commonly associated with negative or positive sentiment, which are given an empirically determined weight; new texts are classified by summing the weights of the words they contain. Thelwall *et al.* [19] report accuracy on Twitter data of 63.7% for positive sentiment and 67.8% for negative when predicting ratings on a 1-5 scale, and accuracies near 95% when predicting a simple binary positive/negative label. However, even though their word lists and weightings are determined for OSN data (including Twitter), this approach may suffer when faced with social text with new words, unexpected spellings and context-dependent language and meaning (see [12]).

For a ML-based option we used the Chatterbox Sentiment Analysis API,[6] based on statistical machine learning over large, distantly labelled datasets [17]. This data-based approach means it might be expected to handle slang, errorful or abbreviated text better. Purver & Battersby [17] report accuracies approaching 80% using a similar technique on smaller datasets; Chatterbox report 83.4% accuracy in an independent study.[7]



(a) Promoted Tweets

(b) Promoted Trends

Figure 3: Distribution of tweet volumes over time

To compare the two approaches, 100 random tweets were selected from the database and manually labelled for positive or negative sentiment, and both tools were tested on the resulting set. Results showed accuracy of 63% for the lexicon-based SentiStrength approach, compared to 84% for the ML-based Chatterbox approach. Error analysis showed one significant source of this difference to be sentiment expressed in hashtags (e.g. the negative `#shambles`), which were detected better by the ML-based approach, presumaby due to their absence from SentiStrength's predetermined lexicon. We therefore use Chatterbox for our experiments.

5. RESULTS

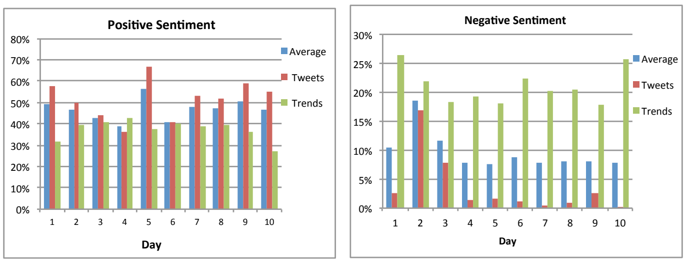
Response volume over time

To examine the spread of engagement for each promoted item over the 10 day period, we analysed the volume of unique tweets each day in response to each promoted item, then averaged the results across all brands. Figure 3 displays the distribution of this volume in response to *promoted tweets* (3a) and *promoted trends* (3b) per brand. On average, promoted trends led to much higher response volumes. However, the highest percentages of *mentions* within responses were from promoted tweets, where an average of 18% of tweets each day included an '@' mention to the brand; promoted trends had an average of only 15% mentions per day. This indicates that for a brand to successfully engage users in the content of the promoted item, a promoted tweet is better for this purpose.

Results confirmed that the greatest percentage of engagement for a brand's promoted item takes place on the first day of promotion. On average, 24% of engagements around the promoted item take place on the first day. The effect is most pronounced for *promoted trends*, with 34% of engagement on average on the first day of promotion, after which the engagement falls dra-

[6] `http://mashape.com/chatterbox-co/sentiment-analysis-for-social-media`

[7] See `http://content.chatterbox.co/Sentiment\%20Analysis\%20Case\%20Study\%20-\%20Chatterbox\%20and\%20IDL.pdf`.



(a) Positive sentiment

(b) Negative sentiment

Figure 4: Sentiment distribution over time.

matically by an average of 25% to 9% by day two and continues to fall thereafter, even if the item is promoted for several days. For *promoted tweets*, the effect is less pronounced: 19% of the engagement takes place on the first day of promotion, with engagement decreasing by 8% by the second day of promotion. However, it does not continue on a steady decline thereafter, but it rises and falls over the next 8 days, although never again reaching the peak of the first day of promotion. This could be due to the fact that a promoted tweet is usually promoted for several days on Twitter where it occasionally appears at the top of different user's timeline were users are repeatedly exposed to the item. This finding can be said to conform to Romero *et al.*'s theory of *repeated exposure* [18].[8]

In general, though, these results show that adoption of a promoted item is not a slow gradual shift over several days (as might be assumed) but rather an immediate incline when exposure to the item is new to users.

Effects on user sentiment

Figure 4 shows the distribution of sentiment in this response traffic over time. On average, positive sentiment outweighs negative sentiment; on the first day, 49% of the tweets were positive. In general, *promoted tweets* lead to more positive sentiment and less negative sentiment than *promoted trends.*

In total, 47% of tweets relating to a *promoted tweet* are positive in sentiment. Day one received the highest percentage of positive sentiment tweets (58%); positive sentiment then continues to dominate over the 10 day period, never falling below 36% of the tweets. Examining *promoted trends*, we found that, on average, only 37% of tweets relating to a promoted trend contained a positive sentiment. On the first day of promotion, 26% of tweets expressed a negative sentiment, 32% expressed a positive sentiment and 42% expressed no sentiment at all. This shows that Twitter users do not tweet as positively about a promoted trend as they would about a promoted tweet. Instead, a large proportion of tweets relating to a promoted trend contained no emotional words, or if they did, the positive and negative sentiments balanced each other out. They generally contained just the promoted hashtag or generally had an objective, matter-of-fact tone (e.g., - “`Get 3G where I live... #O2WhatWouldYouDo`”).

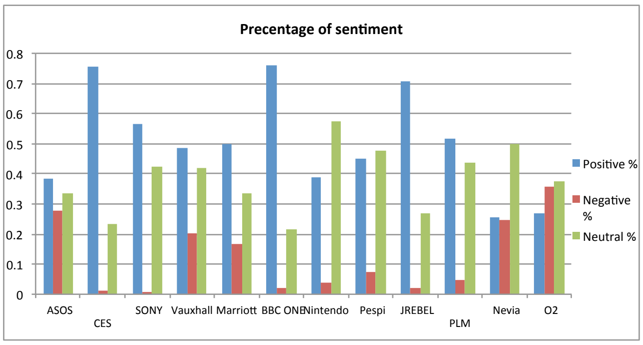


Figure 5: Sentiment analysis by brand

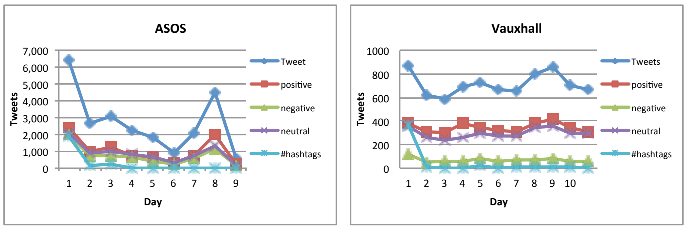


Figure 6: Hashtag related engagements

The sentiment breakdown for each promoted brand item can be observed in Figure 5. We observe that in most cases, the percentage of positive sentiment outperformed that of negative and neutral for promoted items. On average, across all brands (promoted tweets and trends), the average percentage of tweets and retweets[9] which contained a positive sentiment is 50%, that which contained a negative sentiment is 12%, and 38% of tweets had a neutral tone.

Taken together with the analysis of engagement volume, these results show that when an item is promoted, the brand and the item get adopted immediately and regarded quite positively by the engaged users. Twitter users welcome the promoted item on Twitter, which has a positive effect on the tweets expressed. The engagement level reduces to an average of 10% of the total tweets on day two, when the item is no longer being promoted, or is no longer seen as “new and interesting”. However, on average, the positive sentiment expressed still outperforms that of negative sentiment and neutral sentiment each day.

Effect of hashtags on engagement and sentiment

We then performed two example case studies, using the ASOS and Vauxhall brands, to examine the use of hashtags within promoted items. Figure 6 shows the results.

[8] Also see http://advertising.twitter.com/2013/03/Nielsen-Brand-Effect-for-Twitter-How-Promoted-Tweets-impact-brand-metrics.html

[9] We assume that retweeting users share the same sentiment as the original tweet.

ASOS promoted a trend, #AsosSale, on the 19^{th} and 20^{th} of December to highlight their Boxing Day sale on the 26^{th} of December (day 8 of data collection). Although the *promoted* hashtag was virtually discarded by day two of data collection, we found that user engagement (use of hashtag, mentions and tweets) for the forthcoming sale continued. This trend is also apparent in Vauxhall's tweet volumes for their sale which stated on the 27^{th} of December (day one of promotion), and ended the day after our 10 day data collection period. The engagement for Vauxhall remained at a consistent level throughout the event (see Figures 3b and 6), despite the rapid drop-off in use of the promoted hashtag.

6. CONCLUSIONS & FUTURE DIRECTIONS

In this paper we present a measurement-driven study of the effects of promoted tweets and trends on Twitter on the engagement level of users, using a number of ML and NLP techniques in order to detect relevant tweets and their sentiments. Our results indicate that promoted tweets and trends differ considerably in the form of engagement they produce and the overall sentiment associated with them. We found that promoted trends lead to higher engagement volumes than promoted tweets. However, although promoted tweets obtain less engagement than promoted trends, their engagement forms are often more brand inclusive (more direct mentions); and while engagement volumes drop for both forms of promoted items after the first day, this effect is less pronounced for promoted tweets. We also found that although the volume of tweets is highest in promoted trends, they do not lead to the same level of positive sentiment that promoted tweets do. Hence advertisers should carefully assess the trade-offs between high level of engagement, drop-off rate, direct mentions, and positive user sentiment.

In the next stage of this study we will investigate the effect of individuals' influence on the take-up of promoted tweets and trends by their social graph. We will investigate new data at finer granularity (hourly) for events that are time-sensitive, such as major concert ticket sales. We believe our findings could provide a new insight for social network marketing and advertisements strategies, in addition to comparing different methods of classifying and filtering relevant content.

Acknowledgments

The authors thank Chatterbox for providing access to their sentiment analysis API.

7. REFERENCES

[1] S. Asur and B. A. Huberman. Predicting the future with social media. *CoRR*, abs/1003.5699, 2010.

[2] T. Blake, C. Nosko, and S. Tadelis. Consumer heterogeneity and paid search effectiveness: A large scale field experiment. 2013.

[3] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *in ICWSM 10: Proceedings of international AAAI Conference on Weblogs and Social*, 2010.

[4] T. Y. Chan, C. Wu, and Y. Xie. Measuring the lifetime value of customers acquired from google search advertising. *Marketing Science*, 30(5):837–850, Sept. 2011.

[5] C. M. Cheung and D. R. Thadani. The effectiveness of electronic word-of-mouth communication: A literature analysis. *Proceedings of the 23rd Bled eConference eTrust: Implications for the Individual, Enterprises and Society*, 2010.

[6] A. M. Degeratu, A. Rangaswamy, and J. Wu. Consumer choice behavior in online and traditional supermarkets: The effects of brand name, price, and other search attributes. *International Journal of Research in Marketing*, 17(1):55 – 78, 2000.

[7] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. Master's thesis, Stanford University, 2009.

[8] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: An update. *SIGDKDD Explorations*, 11(1):10–18, 2009.

[9] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, 60(11):2169–2188, 2009.

[10] C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing.* MIT Press, 1999.

[11] Y. Mejova and P. Srinivasan. Crossing media streams with sentiment: Domain adaptation in blogs, reviews and twitter. *Proc. ICWSM*, 2012.

[12] A. Naradhipa and A. Purwarianti. Sentiment classification for indonesian message in social media. In *Cloud Computing and Social Networking (ICCCSN), 2012 International Conference on*, pages 1–5, 2012.

[13] A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the 7th conference on International Language Resources and Evaluation*, 2010.

[14] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135, 2008.

[15] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[16] M. J. Paul and M. Dredze. You are what you tweet: Analyzing twitter for public health. In *Int'l Conference on Weblogs and Social Media, ICWSM*, 2011.

[17] M. Purver and S. Battersby. Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 482–491, Avignon, France, Apr. 2012. Association for Computational Linguistics.

[18] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 695–704, New York, NY, USA, 2011. ACM.

[19] M. Thelwall, K. Buckley, and G. Paltoglou. Sentiment strength detection for the social web. *J. Am. Soc. Inf. Sci. Technol.*, 63(1):163–173, Dec. 2012.