Topic and Sentiment Analysis on OSNs: a Case Study of 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
dacresshana@gmail.com, hamed.haddadi@qmul.ac.uk, m.purver@qmul.ac.uk

ABSTRACT

Social media have substantially altered the way brands and businesses advertise: Online Social Networks provide brands with more versatile and dynamic channels for advertisement than traditional media (e.g., TV and radio). Levels of engagement in such media are usually measured in terms of content adoption (e.g., *likes* and *retweets*) and sentiment, around a given topic. However, sentiment analysis and topic identification are both non-trivial tasks.

In this paper, using data collected from Twitter as a case study, we analyze how engagement and sentiment in promoted content spread over a 10-day period. We find that promoted tweets lead to higher positive sentiment than promoted trends; although promoted trends pay off in response volume. We observe that levels of engagement for the brand and promoted content are highest on the first day of the campaign, and fall considerably thereafter. However, we show that these insights depend on the use of robust machine learning and natural language processing techniques to gather focused, relevant datasets, and to accurately gauge sentiment, rather than relying on the simple keyword- or frequency-based metrics sometimes used in social media research.

I 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 [20]. 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. Understanding the real engagement of the end users with the brands and their OSN presence has given rise to a number of data analytics, sentiment analysis and social media optimisation startups and academic research projects. However, the techniques required pose a number of challenges and pitfalls often ignored by researchers and analysts, and adopting a particular method naively can lead to problems. Significant progress in Natural Language Processing (NLP) and Machine Learning (ML) has produced models for topic modelling designed for social media [23], and high accuracies in sentiment detection (e.g. [26]), even with the possibility of detecting sarcasm [11]; but care still needs to be taken while using and relying on the relevant tools and techniques straight out of the box [8].

In this study we present a focused case study by examining the content and volume of users' brand engagement on OSNs to determine the effect of choice of promotion channel on a brand's influence.[2] 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

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

[2] In this work, engagement is defined as adoption of the content by e.g., replying to a tweet, mentioning the brand name, or including the hashtag in a tweet. We are not able to measure external engagement such as sharing content on other OSNs, or clicking on the links in the tweet.

ISBN: 978-1-62561-000-3

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 observe the need to accurately filter the resulting tweets for topical relevance, and compare simple keyword-based methods with a discrimative Machine Learning (ML) approach. We then classify the tweets by sentiment (positive, negative or neutral), and again compare a range of existing methods and tools. We then use this data to establish the driving factors behind the success of promoted influences and differences between advertising strategies. For both tasks, the choice of classification method makes a significant difference, highlighting the care that must be taken when choosing techniques for this kind of analysis.

The rest of the paper is organized as follows: In Section II we present the some recent related studies. In Section III we describe our case study, dataset and its characteristics. In Section IV we briefly discuss our sentiment analysis and text classification methodology and the challenges which only become apparent upon thorough manual inspection of the data. Section V presents our results and the insights gained from our analysis. We conclude the paper and present potential future directions in sentiment and content analysis in Section VI.

II RELATED WORK

INFLUENCE ON OSNS

Our primary interest in this work is in understanding the factors which govern the effectiveness and influence of campaigns on OSNs. Several recent studies have examined individuals' influence on OSNs [3], and the effectiveness of online advertising [2, 4], but little attention has 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.* [10], 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 such mentions and sentiments across different promotion strategies available to brands on Twitter, thus specifically investigating advertising effectiveness (see Section III).[3]

In a study on the spread of hashtags within Twitter, Romero *et al.* [24] 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 exposure to users who use these hashtags affects the probability that a person will eventually use the hashtag more positively than average. However, they only examined 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, possibly with positive sentiment, from one that fails to attract attention or is associated with mainly negative sentiment? Our study aims to answer this by examining the sentiment and spread of tweets in relation to brands' promoted items.

ANALYSIS METHODS

Sentiment analysis has been approached across many domains, including products, movie reviews and newspaper articles as well as social media (see e.g [18] 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. [14]), but generally depend on large amounts of manual annotation (expensive and often problematic in terms of accuracy) and in some cases the existence of grammatical resources for the language and text domain in question (e.g. [26]). However, some approaches leverage the existence of implicit labelling in the datasets available (*distant supervision*), to avoid the necessity for manual annotation: for example, user ratings provided with movie or product reviews [4, 19]); or author conventions such as emoticons and

[3] Data availability limits us to effects within OSNs; we cannot determine effects on actual clicks or sales.

hashtags on OSNs [7,17,21]). Hybrid approaches also exist, e.g. the use of predefined sentiment dictionaries with weights learned from data (e.g. [27]).

Identifying the topic of text has also received much attention in NLP research, with methods ranging from the use of existing topic resources or ontologies (e.g. [12]) to unsupervised models for discovery of topics (e.g. [23]). The use of machine learning to detect the relevance (or otherwise) of text to a known topic also has a long history, perhaps most well-known in the form of Naïve Bayes filtering for spam filtering [25].

However, research into OSN behaviour or influence sometimes ignores the spread of sophisticated methods available. Sentiment analysis is often performed based on defined dictionaries (e.g. [28]), and topic identification is often ignored, with datasets filtered purely on keywords or simple Boolean queries. Recently, Goncalves *et al.* [8] examined the difference in performance across various sentiment analysis approaches on online text, finding significant variations. The effect of these variations in a specific analysis problem is less clear, though: how much does the variation in sophistication (and accuracy) of these methods actually matter? [22] compared statistical and lexicon-based methods and found significant differences at the level of individual messages, although a correlation at the level of their intended analysis (user profiles). Here, we investigate the effect when considering individual advertising campaigns (promoted items). For text relevance, we compare the use of keywords to Naïve Bayes classification via Weka [9]. For sentiment analysis, we examine three existing and freely available tools: the widely-used Data Science Toolkit's `text2sentiment`[4] based on a sentiment lexicon [16]; the lexicon-based but data-driven hybrid SentiStrength [27]; and a statistical machine-learning-based approach, Chatterbox's Sentimental[5] (see [21]).

III DATA COLLECTION

We set up a crawler to use the Twitter Streaming API[6] 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.

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
Health Care	Promoted tweet	PatientsLikeMe
	Promoted trend	NiveaUK
Retail	*Promoted trend*	ASOS
	Promoted trend	PepsiMax
	Promoted tweet	JRebel
Telecomms	*Promoted trend*	O2 Network

Table 1: Industry sectors and sample brands

IDENTIFYING PROMOTED BRANDS

Twitter distinguishes promoted tweets and trends by the use of a *Promoted* tag. We collected tweets from 11 brands with an active advertising campaign during our study period, across different industry domains, ranging from entertainment to health-care. For each promoted item, the brand names 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 industries' promoted items for 10 day periods between 17^{th} December 2012 and 7^{th} January 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

[4] `http://www.datasciencetoolkit.org/`

[5] `http://sentimental.co/`

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

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 or hashtag. When a brand contained more than one directly relevant hashtag, e.g., `#Coke` and `#CocaCola`, we included all the relevant hashtags.

Twitter users do often repeat their tweets to benefit from repeated exposure. However, 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 during our measurement periods, 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. After manual inspection of many tweets and accounts, we are confident that nearly all spam has been removed from our dataset.

IV TEXT PROCESSING & CLASSIFICATION

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

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 III. However, whilst sampling tweets for spammers, a general problem surfaced. We found that a keyword-based approach tends to be too broad 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 [9]. 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. [13]) with various scaling methods,[7] 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).

To determine a suitable training set size, we then

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

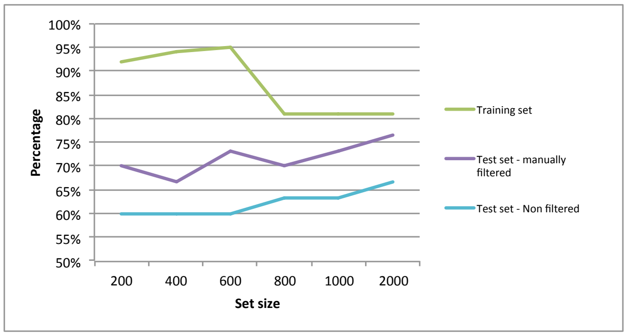

Figure 1: NB accuracy with increasing training data.

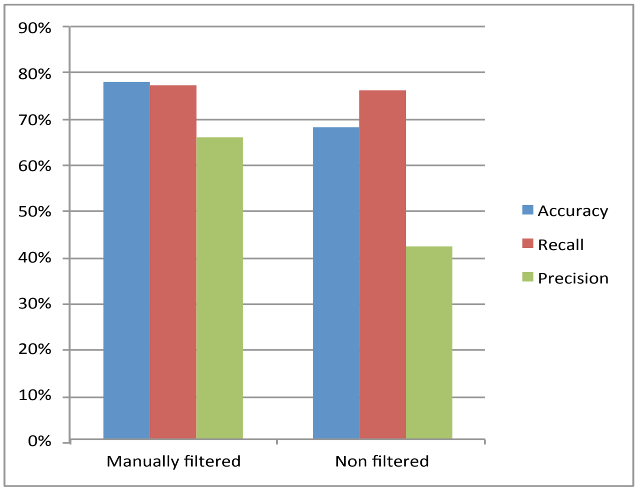

Figure 2: Classification results using Naive Bayes.

varied the training set while testing performance on a held-out test dataset of 30 manually labelled tweets. 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. Figure 2 shows 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. Figure 3 gives details of the per-class predictions: without manual filtering, false positives are more common than false negatives (i.e. too much irrelevant data is slipping through); levels are much closer with filtering.

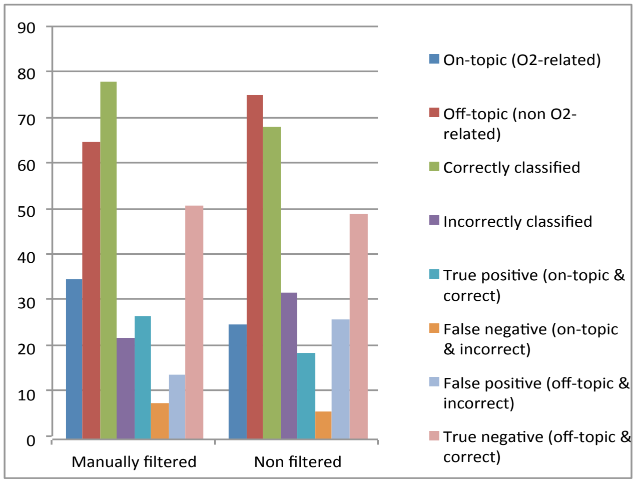

Figure 3: Classification details per class using Naive Bayes.

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 II 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 baseline lexicon-based tool we used the freely available Data Science Toolkit.[8] The sentiment analyser is based on a sentiment lexicon [16]; we therefore anticipate its coverage to be low but take it to be representative of commonly-used lexicon-based approaches.

For a more robust tool for comparison, we examined two alternatives. As a hybrid lexicon/machine-learning tool we chose SentiStrength [27]. This method uses a predetermined list of words commonly associated with negative or positive sentiment, which are given an empirically determined weight (learned from data); new texts are classified by summing the weights of the words they contain. Thelwall *et al.* [27] 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 [15]).

[8] `http://www.datasciencetoolkit.org/`

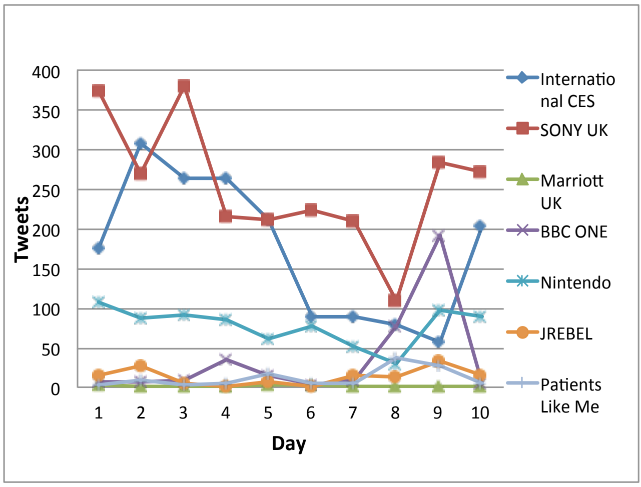

Figure 4: Distribution of promoted tweets volumes over time.

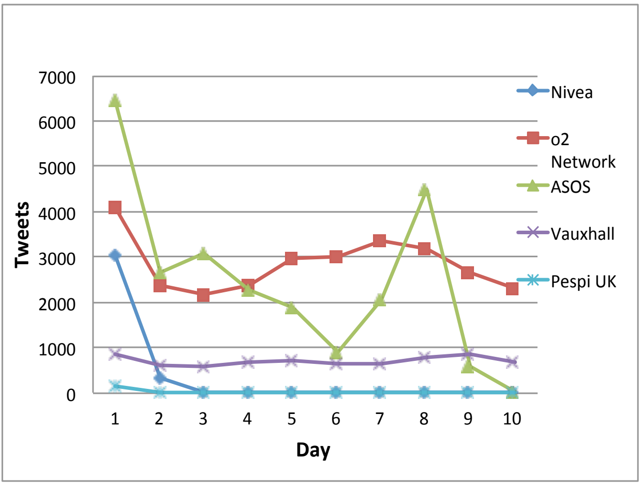

Figure 5: Distribution of promoted trends volumes over time.

For a purely ML-based option we used Chatterbox's Sentimental API,[9] based on statistical machine learning over large, distantly labelled datasets [21]. This data-based approach means it might be expected to handle slang, errorful or abbreviated text better. Purver & Battersby [21] report accuracies approaching 80% using a similar technique on smaller datasets; Chatterbox report 83.4% accuracy in an independent study.[10]

Before applying the sentiment analysis tool, and in order to compare the two approaches, a few hundred random tweets were selected from the database and were read and manually labelled for positive or negative sentiment, and both tools were tested on the resulting set. Results showed accuracy below 50% for the lexicon-based Data Science Toolkit, 63% for the hybrid SentiStrength approach, and 84% for the ML-based Chatterbox approach. Error analysis showed one significant source of the latter 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 as the "robust" tool in our experiments below, and compare to the Data Science Toolkit as a purely lexicon-based baseline.

V 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. Figures 4 and 5 display the distribution of this volume in response to *promoted tweets* (4) and *promoted trends* (5) 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 varies, however: for example, out of the ~30,000 tweets around the O2 Network's promoted trend, 7,965 included an '@'mention to the brand (25%).This indicates that for a brand to engage the maximum number of users, promoted trends are better; but to successfully engage users in a conversation with the brand, promoted tweets may provide a better return.

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 *pro-*

[9] `http://mashape.com/sentimental/sentiment-analysis-for-social-media`

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

moted trends, with 34% of engagement on average on the first day of promotion, after which the engagement falls dramatically 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* [24].[11] They found that repeated exposure to a hashtag within Twitter had a significant marginal effect on the probability of adoption of that hashtag.

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

The sentiment breakdown for each promoted brand item can be observed in Figures 6 and 7, with Figure 6 showing the results obtained using our chosen machine learning method and Figure 7 those obtained using a keyword-based method (see section IV above). We observe that in most cases, the percentage of positive sentiment was higher than that of negative and neutral for promoted items. Notable exceptions are the results for two brands, NiveaUK and O2, where neutral and/or negative levels outweigh positive; the ASOS brand also shows little difference between negative and positive levels. However, comparison of the figures that would have been gained using a keyword-based approach (Figure 7) shows misleading results in precisely these interesting cases: apparent positive levels are higher than negative in all cases. Neutral cases also appear much more common; this is due to the low coverage of the keyword lexicon causing large numbers of results with apparently zero sentiment. Use of the more accurate tool (as objectively assessed – see section IV) therefore does appear crucial.

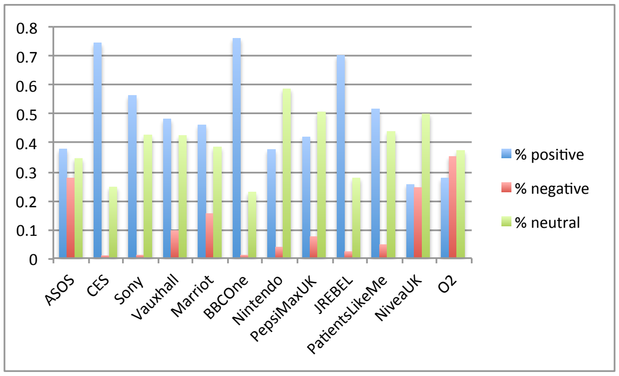

Figure 6: Sentiment analysis by brand - machine learning

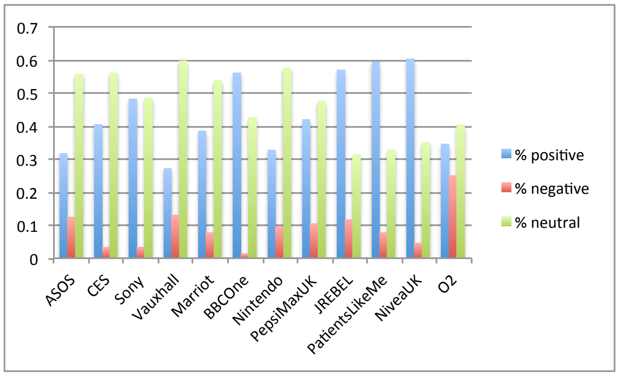

Figure 7: Sentiment analysis by brand - keywords

On average across all tweets and retweets,[12] positivity seems to dominate: 35.8% contained a positive 26.6% a negative sentiment, and 37.5% a neutral tone. This figure is dominated by brands with higher tweet volumes, of course; if we take a macro-average over brands instead, the percentage of positive sentiment is even higher: positive 49.4%, negative 11.1%, and neutral 39.4%.

Figures 8 and 9 then show the distribution of (macro-averaged) positive and negative sentiments 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, 61% of tweets relating to a *promoted tweet* are positive in sentiment. This seems to be influenced to some degree by the original tweet being promoted

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

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

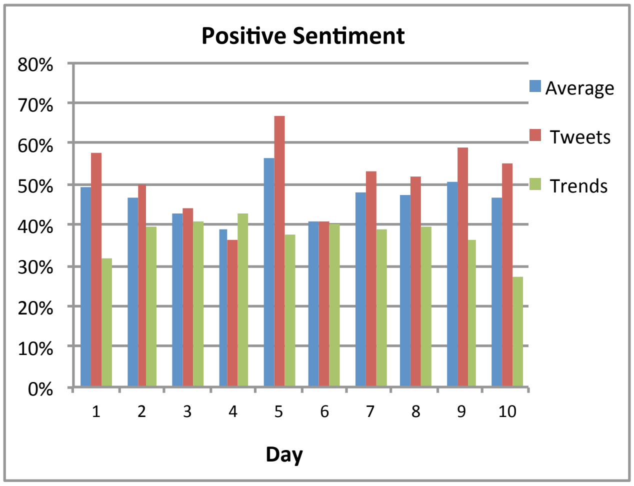

Figure 8: Positive sentiment distribution over time

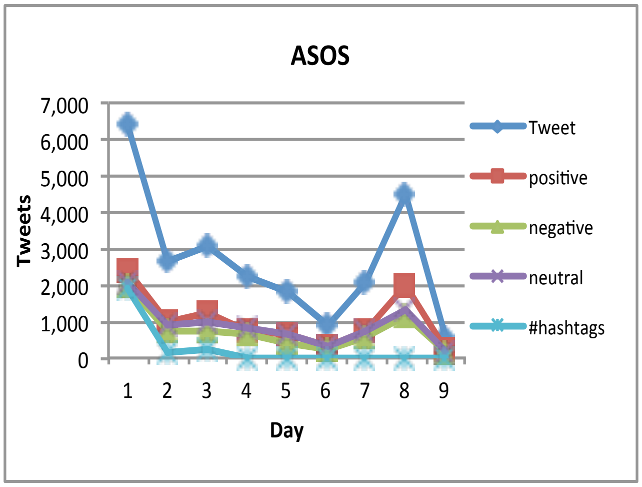

Figure 10: Hashtag related engagements for ASOS.

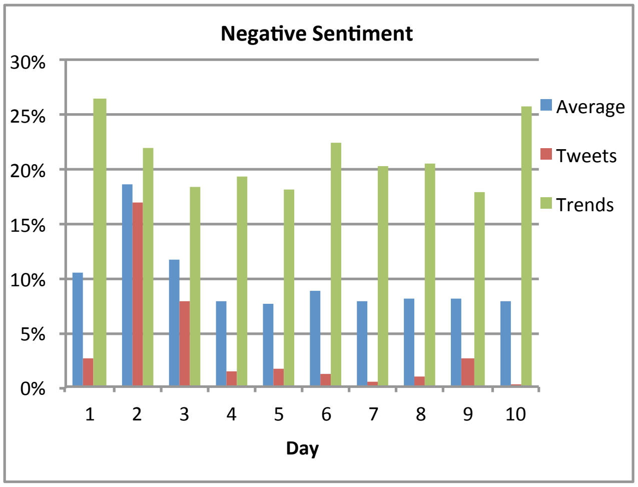

Figure 9: Negative sentiment distribution over time.

(we would expect these generally to be positive), but removing these only reduces the figure slightly, to 57%. Day one received the highest percentage of positive sentiment tweets (72%); 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 34% 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`”).

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. Figures 10 and 11 show the results. 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

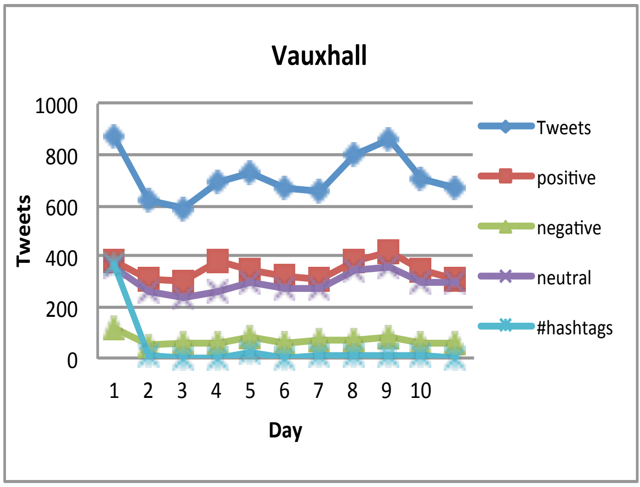

Figure 11: Hashtag related engagements for Vauxhall.

throughout the event (see Figures 5 and 11), despite the rapid drop-off in use of the promoted hashtag.

VI 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 use of accurate methods for sentiment analysis, and robust filtering for topical content, is crucial. Given this, we then see 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 tradeoffs 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 and information flow in the follower/followee graph. We will investigate new data at finer granularity (hourly) for events that are time-sensitive, such as major concert ticket sales. We also wish to perform volume and sentiment comparisons before and after the promotions, examining the characteristics and interaction of new users who had not tweeted about a brand before. The advertising campaigns have very different structure and we need to understand these in details. Promoted trends typically stay on the trends list for a day, and promoted tweets are selectively shown to a subset of users for a period of time selected by the advertiser. Without accounting for such nuances, broad statements on the impact of the two forms of advertising are not conclusive. However in this paper we focussed on insights in using sentiment analysis methods and accurate data labelling. 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 Labs Ltd. for providing unlimited access to the Sentimental API.

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 Õ10: 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] P. Goncalves, M. Araujo, F. Benevenuto, and M. Cha. Comparing and combining sentiment analysis methods. In *Proceedings of the 1st ACM Conference on Online Social Networks (COSN'13)*, 2013.

[9] 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.

[10] 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.

[11] C. Liebrecht, F. Kunneman, and A. Van den Bosch. The perfect solution for detecting sarcasm in tweets #not. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 29–37, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[12] P. Malo, P. Siitari, O. Ahlgren, J. Wallenius, and P. Korhonen. Semantic content filtering with Wikipedia and ontologies. In *IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 518–526. IEEE, 2010.

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

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

[15] 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.

[16] F. Å. Nielsen. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop 'Making Sense of Microposts': Big things come in small packages*, pages 93–98, May 2011.

[17] 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.

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

[19] 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.

[20] 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.

[21] 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.

[22] D. Quercia, J. Ellis, L. Capra, and J. Crowcroft. Tracking "gross community happiness" from tweets. Technical Report RN/11/20, University College London, Nov. 2011.

[23] A. Ritter, C. Cherry, and B. Dolan. Unsupervised modeling of Twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180, Los Angeles, California, June 2010. Association for Computational Linguistics.

[24] 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.

[25] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. A Bayesian approach to filtering junk email. In *AAAI Workshop on Learning for Text Categorization*, Madison, WI, July 1998. AAAI Technical Report WS-98-05.

[26] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013. To appear.

[27] 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.

[28] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. 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*, 2010.