Towards Tracking and Analysing Regional Alcohol Consumption Patterns in the UK through the use of Social Media

Daniel Kershaw
Highwire CDT
Lancaster University
d.kershaw1@lancaster.ac.uk

Matthew Rowe
School of Computing and Communications
Lancaster University
m.rowe@lancaster.ac.uk

Patrick Stacey
Management Science
Lancaster University
p.stacey@lancaster.ac.uk

ABSTRACT

Monitoring rates of alcohol consumption across the UK is a timely problem due to ever-increasing drinking levels [36]. This has led to calls from public services (e.g. police and health services) to assess the effect it is having on people and society. Current research methods that are utilised to assess consumption patterns are costly, time consuming, and do not supply sufficiently detailed results. This is because they look at snapshots of individuals' drinking patterns, which rely on generalised usage patterns, and post consumption recall. In this paper we look into the use of social media such as Twitter (a popular micro blogging site) to monitor the rate of alcohol consumption in regions across the UK by introducing the Social Media Alcohol Index (SMAI). By looking at the variation in term usage, and treating the social network as a spatio-temporal self-reporting sense-network, we aim to discover variation in drinking patterns on both local and national levels within the UK. This study used 31.6 million tweets collected over a 6 week period, and used the Health & Social Care Information Centre (HSCIC) weekly alcohol consumption pattern as a ground truth. High correlations between the ground truth and the computed SMAI (Social Media Alcohol Index) were found on a national and local level, along with the ability to detect variation in consumption on National holidays and celebrations at both local and national levels.

Categories and Subject Descriptors

H.5.m [**Information Systems**]: Information Systems Applications—*Miscellaneous*

Keywords

Twitter, SNS, Keyword Analysis, Alcohol, Trend Detection

WebSci'14, June 23–26, 2014, Bloomington, IN, USA.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2622-3/14/06 ...$15.00.
http://dx.doi.org/10.1145/2615569.2615678 .

1. INTRODUCTION

Alcohol consumption is ingrained in British culture. This is reflected in and reinforced by certain British literature such as Ian Fleming novels, in which James Bond occupies a near alcoholic status [23]. Recently there has been growing concern due to ever-increasing consumption levels. This increase has seen the intake of alcohol rise from 3 litres of pure alcohol per capita in the 1930s to 10 litres per capita in 2006 [35]. The increase in alcohol consumption has been linked to an increase in A&E admittance, anti-social behaviour within town centres [7], and a positive relationship to mortality - the majority of alcohol-related deaths across the world come from injury, liver cirrhosis, poisoning, and malignancy, which contributes to 4% of all fatalities per year [38, 32, 5].

Alcohol-related statistics are compiled from a number of sources in the UK public sector, predominantly from departments within the UK Department of Health (DoH); this includes the NHS (National Health Service). However the collection of data is an expensive and long process, consisting of one-on-one interviews and large-scale surveys. The long lead cycle from data collection to final analysis and release means that the data is only a snapshot of the past and not a current understanding of what is happening. This lack of up-to-date information is at the expense of many services that rely on providing support services in relation to the consumption of alcohol; these include town police forces and A&E departments, which base their staffing levels on historical, out-of-date reports.

The methods currently used to collect the data use quantity-frequency questionnaires (QF), which ask participants to characterise their consumption in averages of drinks over a time period and patterns of beverage-type consumption e.g. how much do you drink in a week, what is the most common drink you consume on a night out? This ignores what different types of beverages may have been consumed at the same time frame, or that all alcohol may have been consumed on one day or a week; these sorts of fine-grained insights can't be determined from QF methods [15]. A more accurate method is the time-line (TL) method - it a gives greater insight into people's consumption patterns by asking for specific drinks consumption over a time-frame e.g. all drinks consumed in the past week. This, has been shown to be a valid method for assessing peoples drinking habits in both problem drinkers and casual drinkers. This is because it allows for a higher quality of data analysis to be applied to the data [37]; though compared to QF it is more expensive to deploy [37].

Comprehensive statistics on alcohol consumption in the UK are compiled by the Health and Social Care Information Centre (HSCIC), who produce the “Statistics on Alcohol” report; this assesses the drinking habits of the adult population (aged 16+) that live in private households in the UK. The questions asked stretch across a week’s drinking habits, including the heaviest drinking day [21]. As mentioned before, the lag between gathering the information to publishing it means that it may quickly become obsolete, and the methods used (QF over TL) may not give the necessary insights for stakeholders. Another issue with reports of this nature is that there are tendencies to underestimate the total consumption of alcohol by up to 40% [5]. This can be seen by extrapolating the number of units consumed per capita from the survey data to the total consumed by the population, compared to real sales figures of alcohol in the UK. This under-reporting has been put down to a number of reasons; selective reporting from people unwilling to report how much they actually drink, recall bias from not remembering what has been drunk as a side-effect from excessive consumption, and accidental under-estimation through mis-estimation of measures [5].

1.1 Research Question

Micro-blogging social networking sites such as Twitter allows users to share up to date information in 140 characters. Twitter currently has 200+ million users with the UK accounting for 32.3 million of them [26]. The reasons that bring users to Twitter can be broken down into a number of key concepts under the umbrella of *frequent brief updates about personal life activities* creating *People-based RSS feeds* [39]; these can be understood as interesting things that happen to people in their day-to-day lives. By keeping up-to-date with such information, users can more readily stay in touch with each other and maintain social relationships, as well as raise their visibility, gather information, seek help, and release emotional stress. These are activities that they may not be able to otherwise accomplish on a day-to-day basis. Accessing and using Twitter can be seen as pervasive and unobtrusive - there are many different mediums one can access the system through (e.g. mobile, computer, smartTV), and the limitation of 140 characters requires minimal effort on the part of the user. This induces users to tweet when they are in a variety of situations, such as: consuming or going to consume alcohol, out at a pub/club/bar, and/or are feeling the effects of alcohol [39, 22].

Given the wide spread uptake of Twitter, this poses the question “*is it possible to characterise and model UK alcohol consumption patterns based on social media data such as Twitter, and if so is there a variation across geographical location in drinking patterns and terminology usage?*” This research differs from previous influenza tracking on SNS’s (Social Networking System) by analysing Geo-located tweets from the UK to detect variations in alcohol consumption patterns in the UK at a regional, regional postcode and at a postcode level for indication of alcohol consumption. At the same time comparisons between different Geo-locations will be used as a comparative insight into different locations’ consumption patterns and language usage.

In this paper we present a method to analyse alcohol-related tweets, how their scores and term frequencies differ across geographical locations and correlate to alcohol consumption patterns. Modelling this data on a ground truth has allowed for the creation of near real-time statistics of alcohol consumption patterns, allowing for the analysis over the long term, as well as on a daily and even hourly basis. This provides a greater insight for services to plan their resource allocation according to new trends.

The contributions of this paper are as follows:

- An approach to model a populations alcohol consumption pattern on Social Media data.
- The discovery of regional variation in relative consumption patterns and term distribution.
- The identification and understanding of how social events effect the overall level of alcohol consumption over an extended period of time.

2. RELATED WORK

There have been many research efforts to utilise social networks and big data to discover real-time information about health-related topics, trends and events. This was first seen in research from Google’s Flu Trend[1] and Yahoo;[2] using their search history logs to look for trends in the variation of frequency of terms associated with influenza like illnesses (ILI), over time. This achieved a high coefficient of determination of 0.4250 [30, 20]. This led to similar research that looked for trends in ILI on social media and micro-blogging sites such as Twitter.[3] A number of approaches were taken; for example, during the swine flu outbreak of 2009-2010, outbreaks were assessed and changes in terminology from “Swine Flu” to “H1N1”’ were analysed [31, 9]. Their research focused on tweets that expressed concern over H1N1 - they found a strong correlation in conjunction with news stories and reports about the outbreaks. Previously with H5N1 (Bird Flu) there had been the move to detect outbreaks in different regions [11] by modeling trends on Twitter using key term models against health service data on influenza like illness (ILI) outbreaks. This returned a high correlation of above 0.80, and was on the way to predicting outbreak through weighed terminology and semi-dynamic key term sets [24]. This model of monitoring a small number of keywords has also been shown to work for estimating alcohol sales in the USA [12]. It showed that the selection of the keywords can be sometimes be problematic, but a combination of “drunk”, “hungover” or “hangover” produced a high estimate of sales of alcohol in the USA, especially when a seven day lag was added to account for drinking the alcohol in the following week.

Analyzing social media content is a growing field of research. One of the main sources of data are sites like Twitter in the West and Weibo[4] in Asia. Initial research looked into topic discovery and which topics were trending [28]. This form of research also performs sentiment analysis of tweets for stock market predictions of certain companies [4], as well as detecting the location of users’ tweets concerning earthquakes in Japan, in order to warn other cities of impending shockwaves - tweets are created and communicated at a faster rate than the shock waves can travel [33].

[1] Google Flu Trend, http://www.google.org/flutrends/

[2] Yahoo, http://www.yahoo.com/

[3] Twitter, http://www.Twitter.com/

[4] Weibo, http://www.weibo.com/

Geographical locations have been used in the past to model topics and to some extent language on Twitter. Attempts were made to model Twitter users' locations by looking at their tweet history for terms that have a higher Geo-location weight, e.g. 'Purdue' would place the tweet in Indiana. This achieved a 51% accuracy of placing users within 100 miles of their 'home', however this did not take into account diachronic and synchronic differences in users vocabulary [8]. Twitter topics were also modelled by looking for variation between language in topics across locations, but this discovered only moderate differences e.g. the sports teams people were supporting [18]; this is because it was only assessing key terms and not the structure of the tweet itself. This bears similarities to the methods used to measure people's life satisfaction on Twitter by modelling Geo-located tweets with an LDA model in order to compare regions of the USA. Regional correlations were found between tweets about 'disengagement' and lower life expectancy, as well as between 'money' / 'work' and being more "well off" [34]

The combination of time and location has also been used for measuring and monitoring depression among Twitter users [14]. The authors initially formed a ground truth by monitoring known people with depression on Twitter, extracting a number of indicators from their activities online, allowing them to class tweets as "depression inductive". These features were then used to predict which tweets where highly "depression indusive". These models were used over Twitter data sets for different time-frames and granulates, allowing for comparison of states through a measure of Social Media Dispersion Index (SMDI). This showed a high correlation with national data on depression, identifying Detroit as the most depressed and Portland as the least.

3. DATA COLLECTION

For the experiment, that we will describe bellow, we collected six weeks worth of tweets in the time period 27th November 2013 til 9th January 2014. The Twitter public Streaming API[5] was utilised; this allowed for a bounding box to be placed around the UK, only allowing tweets Geotagged to the UK to be mined; in total 31.6 million tweets were collected. During this ten week period there was the Christmas Holiday Period, which also included New Years Eve.

Data from the Health & Social Care Information Centre (HSCIC) was used as the ground truth to test the model on; this came from the 2011 report that showed the last day on which a person binge-drank [21] in the week of the survey. Binge drinking in the UK is defined as drinking twice the recommended units of Alcohol within 24hrs; for a man this would be three strong beers (8 units), and a woman would be 2 large glasses of wine (6 units) [29]. This was chosen as it was the only data available with granularity to a day.

4. METHOD

In this section we will introduce the method which will be used to track mention of alcohol terms in tweets. The complete set of tweets is denoted T with a single tweet defined as t such $t \in T$. However we will need to define the subset of tweets grouped by days, hour in days, and location in hours in days. This will then use the granularity of tweet sets to compare against the ground truth.

To group the tweets by day we created a function $day(t)$ which returns the day the tweet was created on, this is then used to identify the day a tweet was posted:

$$T_k^D = \{t : day(t) = k, t \in T\} \tag{1}$$

$$k \in [1, 2, ..., 42] \tag{2}$$

Where k is the number of days since 27th November 2013.

To group by hour's in a day first we group by the day then by the hour creating a function $hour(t)$ which returns back the hour within the day that the tweet was created on:

$$T_{lk}^H = \left\{t : hour(t) = l, t \in T_k^D\right\} \tag{3}$$

$$l \in [0, 1, ..., 23] \tag{4}$$

Where l is the hour within the day.

To group by location in a given hour we first group to the hour then use the function $location(t)$ which returns the location the tweet was created in.

$$T_{lkp}^P = \left\{t : location(t) = p, t \in T_{lk}^H\right\} \tag{5}$$

$$p \in P \tag{6}$$

Where, P is the set of all the postcodes within the UK.

The function location(t) is used again to find the subset of locations in a subset of tweets in a day.

$$T_{kp}^P = \left\{t : location(t) = p, t \in T_k^D\right\} \tag{7}$$

Key-terms (markers) (Table 1) that denote alcohol consumption are defined as $m \in M$. A tweet t is defined as a bag of words $w \in W$. The list of words defined is returned by a function of a tweet $tokens(t)$; this will return a list of words, including duplicates.

If a key-term m appears in a tweet t then it is marked 1 else it is given 0.

$$c(t, m) = \sum_{w \in tokens(t)} f(w, m) \tag{8}$$

$$f : W \times M \to \{0, 1\} \tag{9}$$

The sum of all key-terms is taken from $tokens(t)$ and divided by the number of tokens from the key-term list used.

$$s(t, M) = \frac{\sum_{m \in M} c(t, m)}{|tokens(t)|} \tag{10}$$

This would then mean that an SMAI (Social Media Alcohol Score) for a given set of tweets is the average of all the scores.

$$\text{SMAI}(T, M) = \frac{\sum_{t \in T} s(t, M)}{|T|} \tag{11}$$

The model which has been used was based on an influenza-like illness (ILI) detection model for Twitter [24]. It has been modified as the original model placed more relevance on completeness of the keyword set, thus would have a score higher than one. We have modified it so that we can detect a signal which has a score of between 0 and 1, thereby giving the alcohol signal strength of a given day or hour.

[5] Twitter Streaming API, `https://dev.Twitter.com/docs/streaming-apis`

drunk	wine	wasted
pissed	hungover	hangover
wine		

Table 1: Keyterms used as markers to indicate alcohol consumption

4.1 Data Processing

The set of tweets (T) are grouped into their respective T^D and T^H groups; then subsets were taken based on the Geo-location of each tweet (t). There are four Geo-location groups; National, Regional, Post Code District and Post Code, e.g. a tweet from Post Code LA1 would appear in the LA1 set, LA set, which is itself part of the North West set, which is in turn part of the national set. The kd-tree data structure in SciPy was used to allow quick nearest-neighbour look-up [27]; this was used to find the shortest distance between a tweet and a central post code.

The whole system was implemented using the map reduce pattern to utilise the parallelisation power of the Hadoop [6] framework. In the map stage the SMAI (Social Media Alcohol Index) was computed for each tweet, which was then mapped onto 8 sets (4 locations sets crossed with 2 time sets) that expands the whole data set from 31.6 million to 252.8 million tweets to be processed. The reduce stage calculated the alcohol score for the set, along with the relative key term probability, and collocations for the corpora of all the tweets in the set which had a score greater than 0.00. Both map and reduce programs were written in Python using the MrJob,[6] this allowed for the NLTK framework [3] to be utilised for text tokenisation, stop word removal and collocation algorithms.

An interactive map[7] of all the Twitter alcohol scores was produced. This shows the Twitter alcohol score output of each Geo-location in the form of a choropleth map, with a time line slider allowing the user to change the data view; giving the ability to see the relative colour changes over time. Zooming to different levels of the map reveal different granularity of the data on a geographical level.

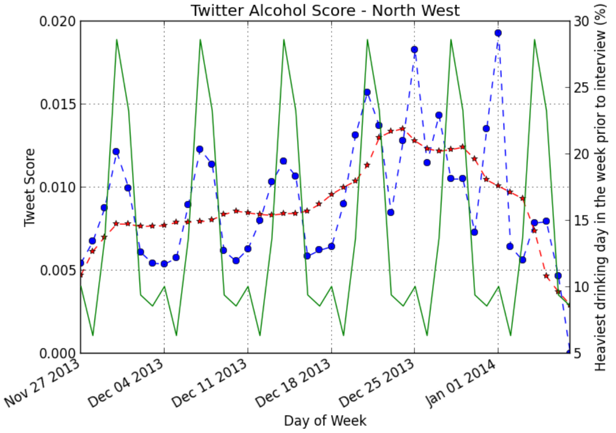


Figure 1: Daily SMAI for whole of UK over 6 week study period. Green line HSCIC Ground Truth data. Blue line daily SMAI. Red Line 7 point moving average.

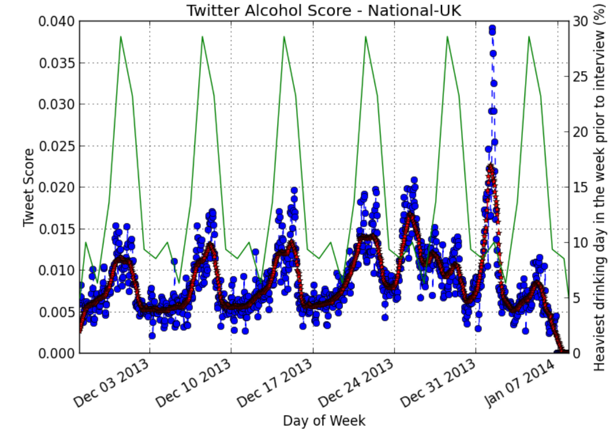


Figure 2: Hourly SMAI for whole of UK over 6 week study period. Green line HSCIC Ground Truth data. Blue line daily SMAI. Red Line 24 point moving average.

4.2 Quantitative Analysis

A Pearson's coefficient and its significance probability (p-value) was calculated between each of regional daily SMAI (Social Media Alcohol Index) against the ground truth (HSCIC) data (Table 2). This was done on each given week within the six week time frame of the study; thus to see if the model maintained for a given week. The highest correlations were seen in Wales (South) in week one and Yorkshire Humber, East England and Scotland (South Central) in week two with the correlations of 0.97, with low p-values. This showed that the model holds up across the regional and national Twitter sets. However the correlation dropped for each consecutive week; this can be attributed to the period that the Twitter data was collected over as it overlapped on the winter holidays within the UK. This is a period which included Christmas and New Year celebrations, both of which are known for people socialising more than normal leading to a 41% increase in alcohol consumption [1] - this means that the UK is the highest alcohol consuming G7 nation for that period [2]. This can be seen through the model not maintaining the correlation where the 7 point moving average is increasing on all SMAI graphs at all geographical levels (Figure 1).

As well as computing the SMAI for each set the relative frequency of each term was assessed. This allowed the assessment and comparison of the terms independent of each other. Initially a cross Pearson's correlation (Table 3) was calculated across the term probability for each region; this indicated that each of the regions had a relative similar distribution of relative term usage. Though slight differences can be seen in Northern Ireland and Wales (North) where 10 and 11 correlations where less than 0.9.

Across the board results from the Channel Islands are very low or 0.00 (Table 3). This can be explained by looking at the master Twitter set, which indicated that there were significantly fewer tweets for that region compared to others, this meant that for days key terms may not have

[6] MrJob, `https://github.com/Yelp/mrjob`

[7] Twitter Alcohol Map, `http://alcohol-Twitter-map.heroku.com`

	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6
National UK	0.93 ***	0.96 ***	0.86 **	0.74 **	-0.23 *	0.05 *
North West	0.92 ***	0.97 ***	0.84 **	0.76 **	-0.22 *	0.11 *
Yorkshire & Humberside	0.93 ***	0.96 ***	0.79 ***	0.71 **	-0.41 *	0.00 *
Greater London	0.86 **	0.93 ***	0.80 **	0.67 **	-0.27 *	0.06 *
South West	0.94 ***	0.94 ***	0.81 ***	0.66 *	-0.33 *	0.05 *
South East	0.91 ***	0.96 ***	0.87 ***	0.58 **	-0.29	0.06 *
Northern Ireland	0.91 ***	0.89 ***	0.80 **	0.57 **	-0.12 *	0.17 *
West Midlands	0.88 ***	0.96 ***	0.84 **	0.59 *	-0.26 *	0.09 *
Channel Islands	0.00 *	0.00 *	-0.30*	-0.35 *	-0.37 *	-0.22 *
Home Counties	0.91 ***	0.95 ***	0.90 ***	0.78 **	-0.24 *	0.06 *
Scotland (North)	0.93 ***	0.96 ***	0.88 ***	0.95 ***	-0.08 *	-0.06 *
East England	0.94 ***	0.97 ***	0.85 ***	0.72 **	-0.20 *	0.052 *
Scotland (South & Central)	0.89 ***	0.97 ***	0.93 ***	0.88 ***	-0.16 *	-0.08 *
Wales (South)	0.97 ***	0.90 ***	0.89 ***	0.78 **	-0.27 *	-0.04 *
Wales (North)	0.96 ***	0.98 ***	0.93 ***	0.76 **	-0.33 *	0.19 *
East Midlands	0.90 ***	0.90 ***	0.69 **	0.69 **	-0.19 *	0.09 *
North East	0.94 ***	0.91 ***	0.79 **	0.81 **	-0.27 *	0.04 *

Table 2: Pearson Correlation of Regional SMAI with NHS Alcohol Data, $* * * = \mathrm{p-value} < 0.01, ** = 0.1 < \mathrm{p-value} > 0.01, * = \mathrm{p-value} > 0.1$

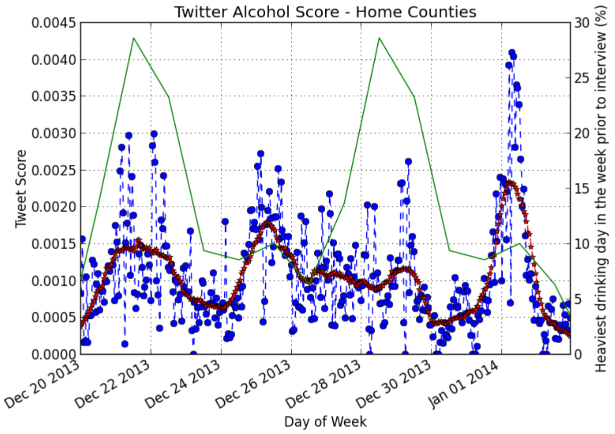


Figure 3: Hourly SMAI for Home Counties over the 2 week Holiday Period. Green line HSCIC Ground Truth data. Blue line hourly SMAI. Red Line 24 point moving average.

been used. This could come from a number of factors; low population density, limited demographics using Twitter, or different culture compared to the rest of the UK, or that the Channel Islands are closer to France than the UK so may not have been included wholly in the bounding box when gathering the tweets.

One final correlation was made against the ranking of regions on if they drank in a week; this was to see if the SMAI indicated variation in tendency to drink over regions and not just variations in patterns in a region over time (Table 4). The rankings of a tendency to drink were based on the HSCIC's Alcohol Statistics report's percentage of a population within each region that drank in a week [21] - this has been used to rank regions across multiple reports. The average SMAI for each region for each week was taken and correlated against the HSCIC data, a Pearson's coefficient of 0.77 was achieved. One of the issues with this measure though is that the data combines the various Wales and Scotland groups into a combined 'Scotland' and a combined 'Wales', so averages had to be taken to combine the data for the bigger sets.

4.3 Qualitative Analysis

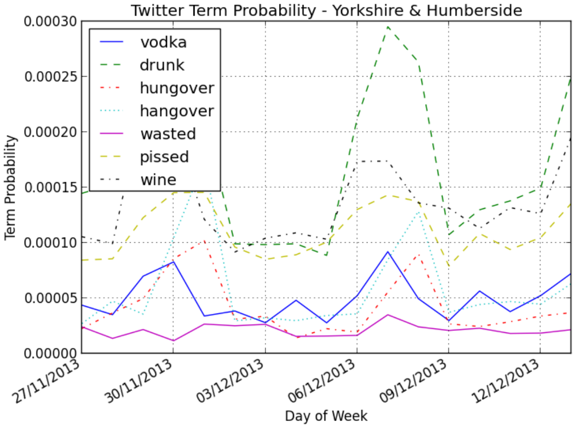


Figure 4: Daily Term Frequency for Yorkshire & Humber 2 weeks starting 27th November

The results indicate that over the winter holidays alcohol consumption increases through the upwards trend of the 7 point moving average on both the hourly and daily charts (Table 1). Although there are lapses in this trend between the two events, this could be due to people going back to work and/or wanting to give their liver a rest.

When looking at a more detailed SMAI graph from the festive holidays there appears to be a trend in increased drinking up to and on Christmas Day (Figure 3), then decreasing afterwards, and spiking upwards again on New Year. There appears to be some interesting increases in SMAI on the weekends either side of Christmas Day - there is a spike in the score; the first one could be an effect of the final day of work and people going out with colleagues to party. Though after Christmas there is a relative plateau (after an initial decrease) which is higher than normal, this could be from people staying off work as Christmas fell mid-week, with reductions only occurring after that weekend before a spike at New Year.

When looking at the hourly graphs from before Christmas a more fine grained understanding can be deduced about potential 'normal habits'. Within the Yorkshire & Humber (Figure 5) on a Saturday from midnight there is a prominent drop in the number of tweets with an increase from Sunday midday, this could be due to people going home and going to sleep, or it may be that they are unable to tweet due to dead barrettes or consuming to much alcohol to tweet.

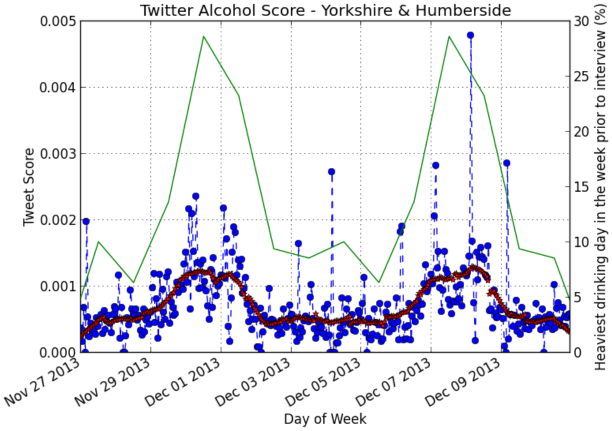


Figure 5: Hourly SMAI for Yorkshire & Humber for initial 2 weeks of study. Green line HSCIC Ground Truth data. Blue line hourly SMAI. Red Line 24 point moving average.

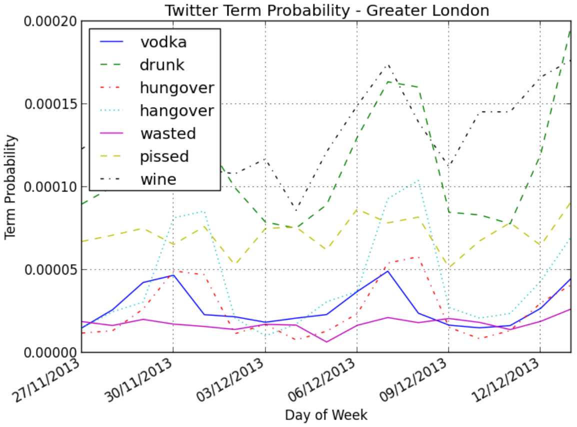


Figure 6: Daily Term Frequency for Greater London 2 weeks starting 27th November.

Though on Sundays a spike in SMAI (Figure 5) can be seen in certain locations, this could come from the change in term distributions as shown in Figure 4 where on the Sunday there is a reduction in the usage of words such as 'vodka', 'wine' and 'drunk', but there is a relative increase in words such as 'hungover' and 'hangover'. Other variations in term frequency can be seen in Greater London (Figure 6) where there is a more prevalent usage of the word 'pissed' which appears to spike on weekends and midweeks, this could be from 'pissed' being used more by students who traditionally go out more on a Wednesday than many other demographics [19].

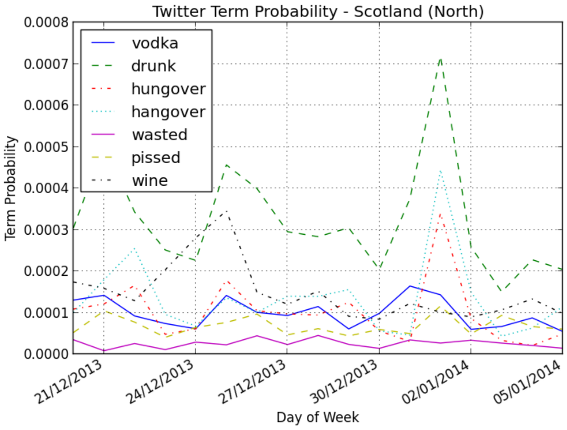
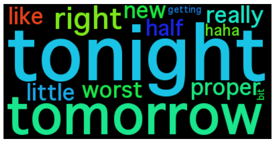


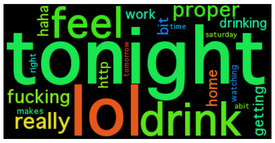
Figure 7: Daily Term Frequency for Scotland (North) 2 weeks starting over Christmas Holiday period.

Around Christmas some terms appear to increase relatively more than others, this can be seen for "wine" which at the time of around Christmas was used more than drunk; this could mean that the drink of choice for around the holidays is Wine and not the other drinks like Vodka (Figure 7). Though generally the terms which were used more are the more prevalent ones in the weeks before.

Some of the characteristics which are seen in the pattern of the probabilities could be down to the concept of the "rich-get-richer phenomenon" where the popularity of already popular items increases faster than less popular items [17]. This was seen in the popularity growth of hash tags on Twitter through a language based study of the spreading of hash tags [13].



(a) Word Cloud for Yorkshire & Humbside for the 6th December 2013 (Friday)



(b) Word Cloud for Yorkshire & Humbside for the 7th December 2013 (Saturday)

Figure 8: Word Cloud for Yorkshire & Humbside

A lot of the collocated words which were coming through were very likely to be talking about the process of getting drunk, such as words “getting” and “drinking” (Figure 9). The aspect of time is also brought in with people reporting that they are going to be drinking in the future by using “tomorrow” or “tonight” (Figure 8). It can also be taken from the data that Red wine appears to be the most Tweeted about wine from Red appearing more than Rose and White across the UK (Figure 10).

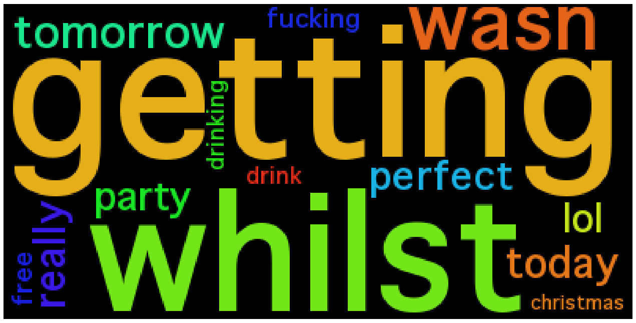


Figure 9: Word cloud on all collocation in Home Counties 21st December 2013

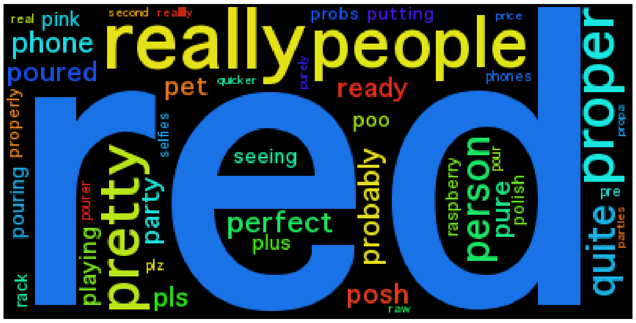


Figure 10: Word cloud on all collocation in North-East

5. CONCLUSIONS AND FUTURE DIRECTIONS

In this paper we presented a method to track variations in drinking throughout regions across the UK by using the social media site Twitter with accuracies as high as 0.97 regionally when compared to the ground truth from HSCIC. The results from this approach could be used in a number of situations from assessing staffing levels in UK A&E departments and policing levels needed in town centres. The method used textual markers as a form of indicator calculating a relative index based on the number of markers in a tweet averaged over a time frame for a Geo-location with data from the HSCIC as ground truth. This shows that there could be many potential benefits of using stream data compares to survey data for time dependent data analysis.

As can be seen through the results above, there is the possibility to model Twitter data against the HSCIC drinking pattern of the nations alcohol consumption. It also shows when nationally and regionally there is a move away from the trend such as national holidays and celebrations. Patterns of lag in terms can be seen across all regions, such as ‘hangover’ spiking 12 to 24 hours after spikes of ‘drunk’. Though there is little word variation between regions in the UK, with the exceptions such as ‘pissed’ in Greater London. This work shows that there is also the need for more granular statistics on people’s consumption patterns, as this was one of the limitations of the work with the ground truth.

However, one of the limitations is that there is a potential population bias. The data for this in the UK is limited, however in the USA 78% of Adults are on-line, although only 17% are on Twitter, and of these the majority belong to a younger demographic [16]. This means that the population on Twitter which is being analysed may not be an exact representation of the population. This has been highlighted by Lazar et al. [25] which critiqued the use of ‘big data’ commenting that systems such as this and Google Flu should be in support of existing systems, and not a supplement.

From this initial exploratory work future work will involve research into predicting drinking patterns in regions in the UK, however this will involve filtering the tweets to remove invalid ones which may skew the results e.i. commercial tweets, and an expansion of markers, potentially removing words which would indicate side-effects and increasing words which would indicate drinking (e.g. alcohol types). Other data sources could be included such as check-ins on Foursquare as an indication of an intention to drink.

Further developments of the system will look for words which increase in popularity over time; this could be used to develop the open key term sets, allowing the system to adapt to the ever changing language in on-line social media. As the research suggested there are slight changes in language in regions, and looking more deeply into this language diffusion could reveal cultural changes in the use of language from how people communicate and express themselves; this would potentially mean moving away from specific term subsets to looking at a region’s whole corpora. This could be seen as timely as language is believed to be fragmenting more and more through the use of on-line media [10].

6. ACKNOWLEDGEMENTS

This work is funded by the Digital Economy programme (RCUK Grant EP/G037582/1), which supports the HighWire Centre for Doctoral Training (`http://highwire.lancaster.ac.uk`).

7. REFERENCES

[1] Christmas statistics - Addaction.

[2] BBC NEWS | World | Europe | UK tops G7 Christmas booze chart, Dec. 2004.

[3] S. Bird. NLTK: The Natural Language Toolkit. *ACL 2006*, 2006.

[4] Bollen, Mao, and Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):8–8, Mar. 2011.

[5] S. Boniface and N. Shelton. How is alcohol consumption affected if we account for under-reporting? A hypothetical scenario. *The European Journal of Public Health*, 23(6):ckt016–1081, Feb. 2013.

[6] D. Borthakur. The hadoop distributed file system: Architecture and design. *Hadoop Project Website*, 11:21, 2007.

[7] R. D. Bromley and A. L. Nelson. Alcohol-related crime and disorder across urban space and time: evidence from a British city. *Geoforum*, 33(2):239–254, 2002.

[8] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: a content-based approach to geo-locating twitter users. pages 759–768, 2010.

[9] C. Chew and G. Eysenbach. Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak. *PLoS ONE*, 5(11):e14118, Nov. 2010.

[10] W. Croft. Social factors in the cultural evolution of language. Comment on "Modeling the cultural evolution of language" by Luc Steels. 8(4):359–360, Dec. 2011.

[11] A. Culotta. Detecting influenza outbreaks by analyzing Twitter messages. *arXiv.org*, July 2010.

[12] A. Culotta. Lightweight methods to estimate influenza rates and alcohol sales volume from Twitter messages. *Language Resources and Evaluation*, 2013.

[13] E. Cunha, G. Magno, G. Comarela, V. Almeida, M. A. Gonçalves, and F. Benevenuto. Analyzing the dynamic evolution of hashtags on Twitter: a language-based approach. In *LSM '11: Proceedings of the Workshop on Languages in Social Media*. Association for Computational Linguistics, June 2011.

[14] M. De Choudhury, S. Counts, and E. Horvitz. Social media as a measurement tool of depression in populations. In *WebSci '13: Proceedings of the 5th Annual ACM Web Science Conference*, pages 47–56, New York, New York, USA, May 2013. ACM Request Permissions.

[15] F. K. Del Boca and J. Darkes. The validity of self-reports of alcohol consumption: state of the science and challenges for research. *Addiction*, 98(s2):1–12, 2003.

[16] M. Duggan and A. Smith. Social Media Update 2013. *pewinternet.org*, Dec. 2013.

[17] D. Easley and J. Kleinberg. *Networks, crowds, and markets*, volume 8. Cambridge Univ Press, 2010.

[18] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. A latent variable model for geographic lexical variation. pages 1277–1287, 2010.

[19] J. S. Gill. REPORTED LEVELS OF ALCOHOL CONSUMPTION AND BINGE DRINKING WITHIN THE UK UNDERGRADUATE STUDENT POPULATION OVER THE LAST 25 YEARS. *Alcohol and Alcoholism*, 37(2):109–120, Mar. 2002.

[20] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2009.

[21] Health and L. S. Social Care Information Centre. Statistics on Alcohol: England, 2012, May 2012.

[22] C. Honey and S. C. Herring. Beyond Microblogging: Conversation and Collaboration via Twitter. pages 1–10, Jan. 2009.

[23] G. Johnson, I. N. Guha, and P. Davies. Were James Bond's drinks shaken because of alcohol induced tremor? *BMJ : British Medical Journal*, 347(dec12 3):f7255–f7255, Dec. 2013.

[24] V. Lampos and N. Cristianini. Tracking the flu pandemic by monitoring the social web. pages 411–416, 2010.

[25] D. Lazer, R. Kennedy, G. King, and A. Vespignani. The Parable of Google Flu: Traps in Big Data Analysis. *Science*, 343(6):1203–1205, Mar. 2014.

[26] I. Lunden. Analyst: Twitter Passed 500M Users In June 2012, 140M Of Them In US; Jakarta 'Biggest Tweeting' City | TechCrunch, July 2012.

[27] S. Maneewongvatana and D. M. Mount. On the Efficiency of Nearest Neighbor Searching with Data Clustered in Lower Dimensions. In *Computational Science — ICCS 2001*, pages 842–851. Springer Berlin Heidelberg, Berlin, Heidelberg, July 2001.

[28] M. Mathioudakis and N. Koudas. Twittermonitor: trend detection over the twitter stream. pages 1155–1158, 2010.

[29] W. Mistral. Binge Drinking: Consumption, Consequences, Causes and Control. *Emerging Perspectives on Substance Misuse*, 2013.

[30] P. M. Polgreen, Y. Chen, D. M. Pennock, F. D. Nelson, and R. A. Weinstein. Using internet searches for influenza surveillance. *Clinical infectious diseases*, 47(11):1443–1448, 2008.

[31] J. Ritterman, M. Osborne, and E. Klein. Using prediction markets and Twitter to predict a swine flu pandemic. *1st international workshop on . . .* , 2009.

[32] R. Room, T. Babor, and J. Rehm. Alcohol and public health. *The lancet*, 365(9458):519–530, 2005.

[33] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *WWW '10: Proceedings of the 19th international conference on World wide web*. ACM, Apr. 2010.

[34] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, and L. Dziurzynski. Characterizing Geographic Variation in Well-Being using Tweets. 2013.

[35] L. Smith and D. R. Foxcroft. Drinking in the UK, 2009.

[36] L. Smith, D. R. Foxcroft, and Joseph Rowntree Foundation. Drinking in the UK, 2009.

[37] L. Strunin. Assessing alcohol consumption: developments from qualitative research methods. *Social science & medicine*, 2001.

[38] World Health Organization. Global Status Report on Alcohol and Health, 2011.

[39] D. Zhao and M. B. Rosson. How and why people Twitter: the role that micro-blogging plays in informal communication at work. pages 243–252, 2009.

	North West	**Yorkshire & Humberside**	**Greater London**	**South West**	**South East**	**Northern Ireland**	**West Midlands**	**Channel Islands**	**Home Counties**	**Scotland (North)**	**East England**	**Scotland (South & Central)**	**Wales (South)**	**Wales (North)**	**East Midlands**	**North East**
North West	1\0	***	***	***	***	***	***	*	***	***	***	***	***	***	***	***
Yorkshire & Humberside	0.96	1\0	***	***	***	***	***	*	***	***	***	***	***	***	***	***
Greater London	0.92	0.94	1\0	***	***	***	***	*	***	***	***	***	***	***	***	***
South West	0.94	0.96	0.94	1\0	***	***	***	*	***	***	***	***	***	***	***	***
South East	0.92	0.95	0.95	0.97	1\0	***	***	*	***	***	***	***	***	***	***	***
Northern Ireland	0.86	0.88	0.83	0.91	0.89	1\0	***	*	***	***	***	***	***	***	***	***
West Midlands	0.96	0.96	0.93	0.96	0.95	0.88	1\0	*	***	***	***	***	***	***	***	***
Channel Islands	0.01	0.02	0.02	0.00	-0.00	0.00	0.01	1\0	*	*	*	*	*	*	*	*
Home Counties	0.93	0.96	0.95	0.97	0.97	0.90	0.96	0.01	1\0	***	***	***	***	***	***	***
Scotland (North)	0.85	0.89	0.86	0.91	0.92	0.90	0.88	0.00	0.91	1\0	***	***	***	***	***	***
East England	0.93	0.96	0.94	0.97	0.98	0.89	0.96	-0.01	0.97	0.91	1\0	***	***	***	***	***
Scotland (South & Central)	0.86	0.90	0.87	0.93	0.93	0.93	0.89	0.01	0.92	0.95	0.91	1\0	***	***	***	***
Wales (South)	0.91	0.92	0.88	0.91	0.91	0.90	0.89	-0.00	0.91	0.90	0.90	0.90	1\0	***	***	***
Wales (North)	0.90	0.89	0.87	0.90	0.88	0.83	0.89	-0.03	0.88	0.84	0.88	0.85	0.89	1\0	***	***
East Midlands	0.94	0.96	0.94	0.97	0.97	0.88	0.97	-0.01	0.96	0.90	0.97	0.90	0.91	0.89	1\0	***
North East	0.93	0.93	0.89	0.93	0.91	0.89	0.91	0.00	0.91	0.90	0.92	0.90	0.94	0.90	0.92	1\0

Table 3: Pearson Cross Correlation or Regional Term Probability Distributions, $*** = p{-}value < 0.01, ** = 0.1 < p{-}value > 0.01, * = p - value > 0.1$

		Average SMAI					
Region	**Drank Last Week (%)**	**Week 1**	**Week 2**	**Week 3**	**Week 4**	**Week 5**	**Week 6**
Yorkshire and the Humber	67	0.00068	0.00066	0.00072	0.00102	0.00111	0.00076
South East	67	0.00063	0.00059	0.00067	0.00087	0.00097	0.00073
South West	66	0.00069	0.00074	0.00076	0.00101	0.00111	0.00079
East of England	66	0.00064	0.00061	0.00070	0.00090	0.00109	0.00075
East Midlands	65	0.00067	0.00072	0.00073	0.00096	0.00108	0.00079
North East	63	0.00072	0.00067	0.00077	0.00103	0.00114	0.00078
North West	63	0.00066	0.00065	0.00069	0.00096	0.00107	0.00075
West Midlands	60	0.00060	0.00059	0.00070	0.00091	0.00101	0.00076
Wales	57	0.00065	0.00069	0.00084	0.00110	0.00114	0.00083
Scotland	56	0.00059	0.00058	0.00066	0.00086	0.00097	0.00079
London	51	0.00048	0.00049	0.00057	0.00065	0.00067	0.00054
Correlation / p-value		0.77 ***	0.60 **	0.36 *	0.52 **	0.67 **	0.47 *

Table 4: Correlations of weekly average SMAI to % people drank in a week, $*** = p - value < 0.01, ** = 0.1 < p - value > 0.01, * = p - value > 0.1$