Assesment of Environmental Stewardship Scheme agreements using the sentiment expressed in trail users' tweets. An exploratory analysis of the Pennine Way National Trail, England.

• Tom Wilson1 • Robin Lovelace1 • Andrew Evans1

**Abstract**

Large and unofficial data sets, often referred to as ‘big data’, are increasingly being used in geographical research and explored as a decision support tool for policy development. Big data has the potential to provide new insight into phenomena about which there is little information from conventional sources. Within this context, the present paper explores the potential of social media datasets to evaluate the aesthetic management of landscape. Specifically, this project utilises the perceptions of visitors to the Pennine Way Trail, which is protected by the UK's Environmental Stewardship Scheme (ESS). The method analyses sentiment in trail users’ public Twitter messages (tweets) with the aim of assessing the extent to which the ESS maintains landscape character within the trail corridor. The method demonstrates the importance of filtering big data to convert the data into useful information. After filtering, the results are based on 161 messages directly related to the trail. Although small, this sample illustrates the potential for social media to be used as cheap and increasingly abundant source of information. We suggest that big data in this context should be seen as a resource that can complement, rather than replace, conventional sources such as questionnaires and interviews. Furthermore, we provide guidance on how social media could be effectively used by conservation bodies, such as Natural England, which are charged with the management of areas of environmental value worldwide.

**Keywords** big data analysis • sentiment analysis • Environmental Stewardship Scheme • Volunteered Geographic Information • National Trails • social media

# Introduction

The Environmental Stewardship Scheme (ESS) is an agri-environmental scheme (AES) in England, integrating environmental concerns into the European Commission’s Common Agricultural Policy (CAP). The ESS provides government-financed payments to farmers and land managers in return for a commitment to farming their land with more care for the environment (Smith *et al.*, 2013). Introduced during 2005 and 2006, the ESS has five primary objectives: Conservation of wildlife and their habitats; the maintenance and enhancement of landscape quality and character; the protection of the historic environment; the protection of soils and reduction of water pollution; and provision of opportunities for people to visit and learn about the countryside (Natural England, 2011). The ESS was established as a ‘broad and shallow’ approach to an AES in order to extend its reach to high proportions of the countryside (Amy *et al.*, 2013). As such it is non-competitive and, at some level, open to all farmers and land managers whose land is part of the farmed environment, and is registered in the Rural Land Register (Natural England, 2013b). The ESS represents the country’s most widespread approach to environmental management, with agreements in place on over 70% of land in the country (Defra, 2013).

England's most stunning and diverse landscapes are connected by the country's National Trail system which comprises of 15 designated National Trails, with a total network length of more than 4000 km. National Trails aim to provide greater access to the English countryside (Long Distance Walkers Association, 2014), rewarding natural adventures, and the opportunity for people to be inspired by varied scenery and landscapes (Wood-Gee, 2008).

Tom Wilson: [[gy10tlw@leeds.ac.uk](mailto:gy10tlw@leeds.ac.uk)](mailto:gy10tlw@leeds.ac.uk)

1 School of Geography, University of Leeds, Leeds, United Kingdom

In 2012 approximately 12 million visits were made to England’s National Trails (Ramblers, 2012). The maintenence and enhancement of landscape quality and character is one of the primary objectives of the ESS, and evidence suggest that the ESS improves biodiversity. Based on these observations, one could assume that the ESS can play an important role in providing a positive experience for visitors to National Trails.

Despite the abundant interactions that National Trail users have with England’s landscapes, and inherently with land managed under ESS, there is currently no method to elicit their opinions regarding the effectiveness of ESS in the maintenance and enhancement of the landscape quality and character. The opinions of trail users are generally limited to broad, large-scale qualitative surveys, such as the ‘Monitor of Engagement with the Natural Environment’ (MENE). MENE examines the adult population’s engagement with the natural environment (Natural England, 2015a). Previous National Trail User Surveys (The Countryside Agency, 2005; Natural England/Countryside Council for Wales, 2007) have not been conducted since 2007, providing a strong incentive to obtain trail users’ opinions using alternative sources.

This research seeks to address this discontinuity in knowledge by exploring the feasibility of utilising the sentiment conveyed within trail users’ public Twitter messages (tweets) to assess the effectiveness of ESS agreements within the corridor surrounding National Trails. Currently, the effectiveness of the ESS tends to be measured in terms of the delivery of environmental benefits and the overall nationwide penetration of the scheme. We aim to devise a process to extract the sentiment conveyed within trail users’ tweets and perform an exploratory analysis to determine whether this information can be used to assess the ESS. The findings of this exploratory analysis have formed the basis of recommendations to Natural England regarding the feasibility of utilising social media data as a method of eliciting trail users’ opinions of ESS and the National Trail System. Based on this evolving policy and research context the objectives of this paper are to:

* Elucidate a process to select tweets that are relavent to the scope of this study from a larger Twitter dataset, and extraction of the sentiment conveyed;
* Detail the spatial analyses of the geographic origins of the tweets, and their viewsheds, to identify patterns in the data;
* Provide broader policy recommendations regarding the feasibility of using social media data as a source of trail user feedback.

This study focuses on a specific National Trail, The Pennine Way National Trail (PWNT). The PWNT travels 431 km (268 miles) along the central upland spine of England connecting the English Midlands to the Scottish Borders (Walk Unlimited, 2014; Long Distance Walkers Association, 2014). The PWNT opened on 24th April 1965 after a 30 year campaign to provide greater access to the English countryside (Long Distance Walkers Association, 2014) and was the first step toward the development of the National Trail system. Along its route the PWNT passes through expanses of land managed under the ESS: 74.08% of the land within 5 km of the PWNT is managed under the ESS. Based on trail-counter data collected between 2004 and 2014, an average of approximately 300,000 people visit the PWNT annually (Natural England, 2014f).

# Background

## Environmental Stewardship Scheme

Agri-environmental schemes (AES) integrate environmental concerns into the European Commission’s Common Agricultural Policy. AES provide government-funded payments to farmers and land managers in return for a commitment to environmentally sensitive farming practices. AES represents the most widespread mechanism of environmental management in England, fundamental to the preservation of the English countryside. The current working implementation of AES in England is the Environmental Stewardship Scheme (ESS).

AES were first introduced to England in the late 1980’s as Environmentally Sensitive Areas, and in the early 1990’s as the Countryside Stewardship Scheme (collectively known herein as ‘legacy AES’). Legacy AES served as the government’s response to increasing levels of agricultural intensification and its negative impacts on wildlife and landscape character (Natural England, 2009b). The aim of legacy AES was to de-incentivise intensive farming (Hodge & Reader, 2010). Although these schemes did reduce intensive farming and mitigate the associated negative impacts, they did little to maintain wildlife habitats and landscape features (Natural England, 2009b) and greater geographic reach was required (Amy *et al.*, 2013). This led to the redevelopment of AES in England, and the introduction of the ESS. The ESS was established as a ‘broad and shallow’ approach –- open to all farmers and land owners -- in order to extend the geographical reach of AES to high proportions of the countryside (Amy *et al.*, 2013). In 2014 the ESS was responsible for the environmental management of approximately 70% of agricultural land in the country (Defra, 2013) and this prevalence means it currently serves as the *de facto* mechanism of environmental protection in England.

The ESS was introduced during 2005 and 2006 and developed as a multi-objective scheme. The ESS aims to provide the funding and guidance to enable farmers and land managers to fulfil the following five main objectives (Natural England, 2011):

* Conservation of wildlife and their habitats
* Maintenance and enhancement of landscape quality and character
* Protection of the historic environment
* Protection of soils and reduction of water pollution
* Providing opportunities for people to visit and learn about the countryside.

There are two main levels to the ESS: Entry Level Stewardship (ELS) and Higher Level Stewardship (HLS). ELS is the underlying broad and shallow approach (Natural England, 2009b) to the ESS which is open to all farmers and land owners in England whose land is part of the farmed environment, and is registered in the Rural Land Register (Natural England, 2013b). HLS is a competitive scheme available to farmers and land managers in pre-defined areas who demonstrate the ability to provide greater environmental benefits. ELS is a stepping-stone to HLS, it is therefore possible for farms to concurrently have ELS and HLS agreements in place. Organic versions of both ELS and HLS (OELS and OHLS respectively) exist. OELS and OHLS follow the same principles as their respective levels but are available only to organic farms or farms in transition between conventional and organic farming. They also offer a premium payment to reflect the inherent environmental benefits delivered through organic farming (Natural England, 2011). Uplands Entry Level Stewardship (UELS) is targeted to severely disadvantaged upland areas. UELS will not be covered in more detail since there is no UELS-managed land within the study region of this research.

### Entry Level Stewardship

Entry Level Stewardship (ELS) is non-competitive scheme with the aims of bringing a large proportion of agricultural land under its management (Hodge & Reader, 2010). ELS provides a straightforward and flexible approach to environmental management. Farmers and land managers can choose from 65 management options. The flexibility of ELS allows environmental management to complement traditional farm operations. Each management option has a pre-assigned point value, for example installing 4 meter buffer strips on cultivated land is worth 400 points per hectare. ELS is adopted by selecting management options that meet the average 30 points per hectare threshold for the whole farm. Adherence to each management option must be sustained for the five year duration of the agreement to receive the bi-annual payments (Natural England, 2011). ELS management options aim to reduce the intensity of farming to improve the environmental quality of the surrounding area. Some options provide incentives to restore and maintain features that are now redundant in production terms, such as hedges and ditches (Hodge & Reader, 2010), that contribute to the character of the landscape (Natural England, 2014c).

The flexibility and ‘broad and shallow’ approach of ELS have led to criticism. Some authors have argued that the flexibility of the scheme jeopardises the environmental benefits provided, and that the extended reach forgoes spatial targeting (Hodge & Reader, 2010). A previous evaluation of the ESS found that many agreements were focused on a limited number of options (Central Science Laboratory, 2007). This suggested that farmers were selecting the management options that would involve the least additional work (Defra & Natural England, 2008) or that could be achieved at zero or minimal cost (Hodge & Reader, 2010). The existence of ‘popular’ and ‘unpopular’ management options has also been discovered, leading to gaps in the provision of environmental benefits. Management options with high point values proved most popular as they enabled farmers to more easily reach their point threshold (Defra & Natural England, 2008).

### Higher Level Stewardship

Higher Level Stewardship (HLS) is a spatially targeted scheme open to agricultural land in specific, pre-designated areas. HLS is a competitive scheme only open to farms that can deliver the greatest level of environmental benefits. In a majority of cases a farm must be part of the ELS scheme prior to applying to HLS. HLS agreements are longer term, seeking to deliver significant environmental benefits over ten years. HLS offers a range of management options that are tailored depending on the specific features of the farm and the environmental management priorities of the surrounding area. HLS payments depend on the specific set of options that are delivered. HLS can also fund capital work projects on features that contribute to the character of the landscape (Natural England, 2014c). Examples include hedging, pond creation, or historical building restoration (Natural England, 2011).

Although HLS also offers some flexibility to farmers and land managers to pick management options to suit their farming practices, it is also a competitive scheme subject to budget constraints. Furthermore, entry into HLS is at the discretion of Natural England, and dependent on the agricultural land in the application being within a targeted area (Quillerou & Fraser, 2010). As a consequence HLS is less affected by criticisms that have been directed at ELS.

### ESS and the National Trail System

England’s network of 15 National Trails provides access to some of the country’s finest countryside (Natural England, 2013). The trails are managed by Natural England, the government agency responsible for the aesthetic (and other) protection of the English countryside. Given the trails’ locations at the heart of the countryside, and the time people may spend on a trail, long distance routes and National Trails provide people with abundant exposure to the countryside and landscapes (Wood-Gee, 2008). ESS agreements within the corridor of a National Trail therefore have a particularly important role to play in providing positive experiences to trail users. As previously mentioned, a primary objective of ESS is the maintenance and enhancement of landscape quality and character. Furthermore, Natural England’s quality standards for the National Trails include enhancement of the landscape, natural, and historic features within the trail corridor (Natural England, 2013a). Previous surveys of trail users of England’s long distance routes (trails of more than 50 miles in length which includes a majority of the National Trails), found that the primary attraction of National Trails is the quality of the scenery and the landscapes through which the trails pass, and that almost 50% of trail users reported that the landscape was the highlight of their visit (The Countryside Agency, 2005; Wood-Gee, 2008).

### Measuring the effectiveness of ESS

Current methods to assess the effectiveness of the ESS generally focus on the environmental benefits provided under the scheme (Franks & Emery, 2013). The report ‘Farming and Nature: Agri-environmental schemes in action’ (Natural England, 2009c) summarises the key achievements of AES to date which include: halting the deterioration of, and restoring of, priority habitats; increasing populations of scarce farmland birds and bumblebee populations; the maintenance and enhancement of landscape character; the protection of historical features; connecting people to the natural environment; and providing a major contribution to climate change mitigation (Natural England, 2009c).

National Trail User Surveys were previously used to capture the opinions of trail users (The Countryside Agency, 2005; Natural England/Countryside Council for Wales, 2007). However these surveys have not been conducted since 2007. As a consequence the only survey of visitors to the countryside is via the ‘Monitor of Engagement with the Natural Environment’ (MENE) that seeks to examine the adult population’s use and enjoyment of the natural environment through a combination of surveys and interviews. The focus of the MENE is on people’s visits to the environment, and the aim is to determine people’s relationship with the natural environment (Natural England, 2015a). However, the MENE is somewhat broad in terms of its definition of the natural environment, which includes “all green open spaces in and around towns and cities as well as the wider countryside... away from home and private gardens” (Natural England, 2015a p1). It is also coarse in that it does not focus on specific environmental spaces such as National Trails. The MENE is a quota-sampled, representative sample of the population that has been conducted annually since 2009 with 46,000 –- 49,000 participants annually (Natural England, 2015a). The annual cost of the MENE is about £400,000 (National Archives, 2014).

As earlier identified, England’s National Trails received 12 million visitors in 2012 (Ramblers, 2012). Based upon ten-year trail counter data, the PWNT receives approximately 300,000 annual visitors (Natural England, 2014f). The absence of a specific survey of trail users means that in despite of the extensive interactions that trail users have with the landscape, the English countryside, and inadvertently ESS agreements in place along the trail corridor, there is currently no mechanism in place to gather and analyse trail user opinions of the ESS or the National Trails.

## Twitter

Twitter is a microblogging site and social web data platform. Twitter allows registered users to publically post short messages of up to 140 characters from a computer or mobile device connected to the internet. These short messages are known as tweets. By default tweets are public, and although it is possible for users to protect their tweets and make them private (boyd *et al*. 2010), a majority of users do not. Users are also able to select whether their tweets include geographic information that denotes from where the tweet originated geographically. This geographic information is often referred to as a geotag, and is a form of volunteered geographic information (VGI; Goodchild, 2007). Current estimates suggest that 1-3% of tweets include a geotag (Morstatter *et al*., 2013; Broniatowski *et al*., 2013; Hecht & Stephens, 2014).

Twitter is also a type of social network whereby users can follow other users and choose to receive and view their tweets. In contrast with other social media sites, however, this follow relationship need not be reciprocal (boyd *et al*., 2010). Twitter user profiles are minimal, compared to those of other social networks such as Facebook or LinkedIn, but are public. A previous study of Twitter categorised the main user intentions of using Twitter as daily chatter about everyday life, conversations between users, reporting and disseminating news (Java *et al*., 2007), and sharing URLs (boyd *et al*., 2010).

As of September 15th 2015, Twitter is the 9th most visited website in the world, and the 11th most visited website in the United Kingdom (Alexra, 2015), excluding mobile users who are not included in Alexra’s data collection (Wilkinson & Thelwall, 2012). In the fourth quarter of 2014, Twitter had 288 million monthly active users (Statista, 2015). The sheer number of active users and the ability for tweets to be posted from anywhere with access to the internet means that Twitter generates a constant stream of information.

Given the large volume of information, hashtags have emerged as a method to label and group topics in Twitter. Prefixing a keyword with a ‘#’ symbol generates a hashtag (boyd *et al*., 2010) which is then included as part of the tweet. The keyword can be anything as chosen by the user, and multiple hashtags can be used. Hashtags are automatically converted to links; clicking hashtags can help others to find tweets of a common theme, or find tweets that are of specific interest to them (Cunha *et al*., 2011).

Twitter provides the functionality to connect to the platform’s Application Programming Interface (API) and stream and gather tweets from the Twitter stream. Several public APIs exist but it is the Streaming API which allows for the acquisition of tweets in real-time. Although the APIs are subject to constant change, Driscoll and Walker (2014) provide a detailed comparison of the Twitter APIs that were available in 2014.

There is a degree of opacity surrounding the Twitter APIs, and none of the public APIs provide direct, unfettered access to Twitter data. The streaming API, for example, is generally believed to be subject to a ‘streaming cap’ of about 1% of all tweets at any point in time (Driscoll and Walker, 2014) and the criteria by which tweets are made accessible to the API, or not, is unknown (boyd & Crawford, 2012).

Nevertheless, the accessibility to a constant stream of user-generated social web data, around 1-3% of which contains VGI, has led to a rapid increase in research based upon social web data. As a consequence there has been development of new approaches to the exploration of phenomena (Wilkinson & Thelwall, 2012). To date, Twitter data has been used to study a diverse variety of topics such as real-time event detection during an earthquake (Sakaki *et al*., 2010), analysis of crisis events such as riots (Proctor *et al*., 2013), public sentiment toward a royal birth (Nguyen *et al*., 2013), and the spread of misinformation and rumour in the wake of a terrorist attack (Starbird *et al*., 2014).

## Sentiment

Sentiment is the view, attitude, or opinion toward an entity such as a situation, event, or item. Knowledge of the sentiment of others is fundamental to the decision-making process and is therefore viewed as a key influencer of action, and central to human behaviour (Lui, 2012). Sentiment is important to individuals and organisations alike. Organisations require knowledge of how consumers and the general public feel out their products and services (Lui, 2012) in order to remain competitive in the market (Nasukawa & Yi, 2003). Individuals seek the opinions of existing users and consumers before committing to the purchase a good or service (Lui, 2012). In sum there is a human desire to know what other people think (Pang & Lee, 2008).

Prior to the growth of the World Wide Web, businesses traditionally obtained the sentiment held by consumers and the general public through the expensive deployment customer satisfaction surveys and the convening of user groups (Nasukawa & Yi, 2012; Lui, 2012). At the individual level, people would simply ask friends and colleagues for their recommendations. Both of these still hold true but the rise of the internet and mobile computing means that people are able to publically review products and services, through dedicated review sites such as Yelp! and TripAdvisor, or via social web sites such as Twitter and Facebook. As a consequence sentiment has established itself as a public commodity and individuals and businesses alike are able to obtain and act upon the opinions and experiences of those they may not personally know, and who are almost certainly not professional reviewers (Pang & Lee, 2008).

### Sentiment Analysis

Sentiment Analysis (SA) is the research area that is concerned with the study of sentiment, opinions, evaluations, attitudes, and emotions towards a subject (Thelwall *et al*., 2012, Lui, 2012), in particular the extraction of the sentiment from text. Due to the significant role that sentiment is deemed to play in influencing human behaviour, SA is a widely researched topic in the field of natural language processing. Initial interest in SA was driven by commercial purposes as businesses sought greater understanding of the role of sentiment in consumers’ decision-making. Much in the same way that the internet, mobile computing, and the social web have shaped the way in which sentiment is publically shared, the technologies have also been responsible for renewed interest in SA, in particular SA of social web data, such as that found on Twitter. For example, SA of social web data has been used in the field of social research (e.g. Thelwall *et al*., 2010; Thelwall *et al*., 2012; Thelwall *et al*., 2011; Go *et al*., 2009), to gain insights into particular events (Thelwall *et al*., 2011), and to study the affective dimension of the social web (Thelwall *et al*., 2012).

In its most basic form SA describes the task of using computer algorithms to automatically identify the semantic polarity of text through identification of the positive or negative opinions that are expressed within that text (Thelwall *et al*., 2010; Pang & Lee, 2008). SA presents a computational challenge since sentiment can be expressed in subtle, nuanced ways, for example in opposing classes (e.g. positive or negative), or with conformity to a numerical scale (e.g. a 1 to 5 star rating) (Pang & Lee, 2008). The process of SA comprises of a minimum of two stages: subjectivity detection and semantic polarity determination. Subjectivity detection determines whether a text, sentence, word, or feature is subjective (i.e. contains sentiment or opinion), or objective (i.e. contains factual information). Semantic polarity establishes whether the identified subjective text is positive, negative, or neutral (i.e. it conveys no sentiment) (Taboada *et al*., 2011). Recent developments mean this two stage model has been extended to also include sentiment strength detection, which measures the strength of sentiment in a text, and multiple sentiment detection, which aims to detect the range of emotions that may be present in a given text (Thelwall, 2013a). SA of social web data presents additional challenges for sentiment analysis algorithms due to poor language use, deliberate non-standard spellings, abbreviations of words, and the use of emoticons (Thelwall *et al*., 2010) and emoji -- all inherent characteristics of social web communications. Features such as these may be misinterpreted or missed altogether by sentiment analysis classifiers that have been designed specifically for commercial purposes (Thelwall *et al*., 2012). Finding a method to successfully extract the sentiment conveyed with social web data is important in this research because a significant task is to extract the sentiment conveyed within the tweets of PWNT trail users.

### SentiStrength

With consideration for the aforementioned challenges associated with SA of social web communications, we will use SentiStrength, a tool that has been specifically developed for the SA of short, informal social web text. SentiStrength is a computer algorithm that uses lexicons (dictionaries) annotated with the semantic orientation of words to calculate the semantic orientation of text (for an example see Turney, 2002) (Taboada *et al*., 2011, Thelwall *et al*., 2012). SentiStrength uses widely available lexicons including the Linguistic Inquiry and Word Count program (Pennebaker *et al*., 2003), the General Inquiry list of sentiment terms (Stone *et al*., 1966), plus annotations that were made during the testing of the tool (Thelwall, 2013). In addition to these core lexicons, SentiStrength also references lexical lists that optimise it for use with social web data. These include an emoticon list, an idiom list, a booster word list, a repeated punctuation list, and a negating word list. The entries in each of these lists has an associated strength score assigned to it. The booster word list is used to strengthen or weaken sentiment words that follow, the negating word list neutralises sentiment words that follow, and the repeated punctuation list boosts the strength of sentiment words with one or more exclamation points (Thelwall, 2013). All lexicons can be modified by the end user of the tool (Thelwall, 2013).

SentiStrength processes text as follows: The algorithm splits text into unigrams (individual words) and punctuation, and these are then queried against the lexicons (Thelwall, 2013). The algorithm identifies the presence of known sentiment-bearing words and predicts the sentiment of the text based upon the frequency of occurrence of the sentiment-bearing words. SentiStrength is also able to detect the strength of sentiment in text because the lexicon contains human-assigned sentiment strength judgements (Thelwall *et al*., 2012). Since positive and negative sentiments can coexist within a text (Fox, 2008), SentiStrength returns two integers; the strength of the positive sentiment (+1 to +5) and negative sentiment (-1 to -5) conveyed in the text. If +1 and -1 are returned it denotes a lack of overall sentiment, i.e. neutral or objective text. It is possible to calculate the overall polarity of a text through addition of the two integers (Thelwall *et al*., 2013).

SentiStrength has been tested on diverse social web data sets and applied to studies in various domains such as a time-series analysis of sentiment expressed on Twitter (Thelwall *et al*., 2011), a determination of emotional diversity in information dissemination on Twitter (Pfitzner *et al*., 2012), a sentiment analysis of commute-related smartphone applications in California (New Cities Foundation, 2012), an assessment of the sentiment of short informal text written about celebrities in German (Momtazi, 2012), and a large-scale sentiment analysis of Yahoo! Answers (Kucuktunc *et al*., 2012).

# Data and Methods

## Data

This pilot study utilised a variety of datasets in order to select trail users’ tweets, determine the sentiment of trail users’ tweets, and perform viewshed analyses of the tweet origins. Specifically, we used the following datasets:

### The Pennine Way National Trail

A GPX file of the route of the Pennine Way National Trail (Walk Unlimited, 2014) was plotted using R (R Development Core Team, 2008), and both 5 km and 25 km buffers were generated. The 5 km buffer represents the geographical scope of the project and is hereafter referred to as the PWNT corridor. The 25 km buffer was used in later viewshed analyses. Figure 3 illustrates the geographic scope of these buffers.

### Environmental Stewardship Scheme Agreements.

A shapefile of the ESS agreement boundaries in England (Natural England, 2014e) was clipped to the extent of the 25 km and 5km PWNT corridors. Non-spatial data for each individual agreement included the level of the agreement (e.g. ELS, HLS), and details about the farm, the duration of the agreement, etc. Table 1 provides a breakdown of the types of ESS agreements in place, and the prevalence of ESS in both the 5 km and 25 km PWNT corridors. Figure 3 illustrates the spatial distribution of ESS agreements within the PWNT corridors.

### Twitter data.

The Twitter data for this research was acquired through a tweet-harvesting project conducted at the University of Leeds (Lovelace, 2014). The dataset was provided in the form of a comma separated file with each row of the file representing a single instance of a tweet. Aside from the text of the tweet (TweetText), the dataset also included additional metadata about each tweet:

* a unique id (TweetID)
* date the tweet was created (DateCreated)
* time the tweet was created (TimeCreated)
* the number of followers of the sender (n\_followers)
* the number of others the sender follows (n\_following)
* the total number of tweets sent by the sender (n\_tweets)
* the sender’s location (user\_location). This refers to the sender’s self-disclosed location from their profile, not the user’s location at the time of sending the tweet.
* The location from which the tweet originated was provided by the geotag fields in the dataset; longitude and latitude. Every tweet in the dataset included this geocoded information, which represents approximately 1-3% of all tweets (Morstatter *et al*., 2013; Broniatowski *et al*., 2013; Hecht and Stephens, 2014).

The Twitter dataset represented 52 days of data collection between 2014-06-03 and 2014-07-25 inclusive and contained a total of 60,466 geotagged tweets and their associated metadata. Figure 4 illustrates the spatial distribution of the tweets within the Twitter dataset.

### Digital Elevation Data

90 m Shuttle Radar Thematic Mapper (SRTM) data (Pope, 2009) was used as the digital elevation model (DEM) in this research. Figure 5 is the DEM of the 25km PWNT corridor.

### Land Cover Data

The 2007 Land Cover Map (LCM) (Morton *et al*., 2011) was used to identify land cover within the study region. The 2007 LCM provides land cover information for the United Kingdom, derived from satellite images and digital cartography. The raster version of the LCM has a 25m spatial resolution, with the value of each pixel representing the most likely broad habitat from 23 classes of broad habitat. Figure 6 shows the spatial distribution of land cover classes within the 25km PWNT corridor. Table 2 provides a list of the land cover classes and their extent within the 25km PWNT corridor.

# Methods

In order to extract the sentiment conveyed within trail users’ tweets and determine the effectiveness of ESS agreements required a multi-stage process. An overview of the steps is provided below, with more detailed description following:

* Spatial selection of tweets based on proximity to PWNT
* Lexical selection of tweets using natural language processing
* Data preparation (removing duplicate tweets, spurious characters)
* Input of tweets’ spatial data imported into GIS
* Sentiment Analysis of TweetText using SentiStrength
* Sentiment analysis output combined with tweet’s spatial data
* Viewshed analyses conducted for each overall positive and overall negative tweet. Viewshed analyses included:
* Calculation of the viewshed
* Determination of majority land cover class within the viewshed
* Determination of the ruggedness within the viewshed
* Determination of the presence of ESS agreements within the viewshed.

## Spatial selection of tweets

Only tweets whose origin was within the 5km PWNT corridor were included in further analysis. Using R (R Development Core Team, 2008), the Twitter dataset was spatially clipped to the 5km PWNT corridor.

## Lexical selection of tweets

Identification of tweets relevant to PWNT use was done through lexical selection. Lexical selection involved searching the TweetText of each tweet using case-insensitive regular expression terms. An approach of trial and error was used to ascertain the search terms that returned relevant results. Table 3 is a list of the 20 search terms that were used in the final selection.

The results for each of the selections were combined into a single dataset. Since some tweets contained multiple search terms, for example “Hiking Malham Cove on the Pennine Way” the dataset was searched and purged of duplicated tweets so they would not be processed further and introduce bias into the results.

Additional data ‘cleaning’ was necessary to remove tweets which, although originated from within the geographical scope of the project, and contained relevant search terms, were not relevant to the project. These included traffic reports, other clearly broadcasted messages, and direct (Twitter-user to Twitter-user) messages. Direct messages are broadcasted tweets sent between specific users, generally in reply to one another, and are identifiable as they include an ['@username'](mailto:'@username').

## TweetText processing

The tweet selection process and data ‘cleaning’ resulted in 161 individual tweets, herein referred to as 'trail users tweets'. The trail users' tweets were those deemed relevant to the project and would be subject to sentiment analysis. As previously described, SentiStrength was selected as the sentiment analysis tool because it is designed for the sentiment analysis of short, informal text that may include abbreviations and slang (Thelwall *et al*., 2011). Prior to SA some additional processing was required to prepare the data. Although SentiStength includes an emoticon list with emoticon polarities to identify the sentiment of emoticons (Thelwall et al., 2012), it is not able to interpret emoji, which are unicode characters used to coney sentiment (Unicode Inc, 2012). A visual analysis of the trail users' tweets revealed that several tweets contained Unicode character combinations. These Unicode symbol combinations were referenced using online tools (Emojipedia, 2015) to find the meaning. The Unicode was then replaced with the text-equivalent sentiment of the emoji so as to preserve the sentiment conveyed. Finally, spurious characters and excessive whitespace, which may have been initially present or introduced by the processing, were removed.

## Sentiment Analysis

SentiStrength accepts tab-delimited text files as input. A tab-delimited text file of the TweetID and TweetText of the processed 161 trail users’ tweets were input. SentiStrength produced a tab-delimited text file with positive (+1 to +5) and negative (-1 to -5) sentiment scores appended to the end of each tweet. This score is the sentiment conveyed within each tweet. A score of +1 or -1 respectively denotes that positive or negative sentiment was not detected. Therefore a tweet with a +1 and -1 score would be treated as neutral (no sentiment conveyed). Although sentiment strength was provided it would not be used here. Rather just the presence of positive, negative, or neutral sentiment was of interest. The sum of the positive sentiment conveyed and negative sentiment conveyed scores was equal to the overall sentiment of the tweet. A positive score denoted positive sentiment and a negative score denoted negative sentiment. A score of 0 denoted no sentiment (neutral)

SentiStrength also provided a duplicate of the TweetText field showing the sentiment score of each individual word. It was noted that two locations on the PWNT; ‘Cross Fell’ and ‘High Force’ were returning negative sentiment scores. This is an example of domain specificity -- i.e. the names have additional meanings (Thelwall *et al*., 2012; Thelwall *et al*., 2011). To override this, both ‘Cross Fell’ and ‘High force’ were added into the idiom list within SentiStrength and given a sentiment score of 0 (no sentiment). The sentiment analysis was conducted a second time after these changes.

The sentiment analysis output was added to ArcMap 10 (ESRI, 2011) and joined by TweetID to the shapefile of tweets for further analysis.

## Viewshed Analyses

Viewshed analysis describes the computational process of predicting the total area that is visible from a point in space (Kim *et al*., 2004). Viewshed analysis has a variety of application including, for example, planning the locations of communication towers (De Floriani *et al*., 1994) and wind turbines (Kinder & Sparkes, 1999), and to identify the impacts of human features on wilderness character in the National Parks of the United States (Tricker *et al*., 2012; Tricker *et al*., 2013).

In this study we used viewshed analysis to determine the visible areas from each of the overall positive, and overall negative tweet locations. Previous research has found that the appeal of National Trails lies within the quality of the scenery and the landscape (Wood-Gee, 2008). Therefore, the purpose of the viewshed analyses is to establish the experiential qualities of the landscape within the viewshed of the sentiment-bearing tweet locations, and to ascertain whether certain characteristics within the viewshed are consistent with a tweet conveying positive or negative sentiment.

Attributes which characterise wildness qualities have been established in previous studies (e.g. Lesslie *et al*., 1993; Carver, 1996; Carver *et al*., 2008). Based upon these works two land attributes were identified as being appropriate for this research; the land cover, and the rugged and challenging nature of the terrain (Carver *et al*., 2008). In addition the extent of the viewshed and the extent and type of ESS agreements within the viewshed, were also of interest.

Viewshed analyses were conducted using The Observation Point Tool in ArcMap 10 (ESRI, 2011). The Observation Point Tool allows for the maximum viewshed distance to be set for each point in the dataset. This parameter, RADIUS2, was set to 20 km meaning that the viewshed would only be considered within this distance. This distance was consistent with previous work (Natural England, 2013c), and necessary because viewshed analysis is a computer intensive process and defining the maximum distance would benefit the processing time. The SRTM digital elevation data (Pope, 2009) was the other input required for the viewshed analysis and was clipped to the 25 km PWNT corridor. This was to ensure a potential viewshed of 20 km for each tweet within 5 km of the trail.

The Observation Point Tool in ArcMap 10 (ESRI, 2011) allows for a maximum of 16 points to be analysed at one time, and the output can be used to determine which cells are visible from each point. For this study, however, the viewshed for each tweet was calculated independently using the ModelBuilder facility within ArcMap 10 (ESRI, 2011). The model iterated through each of the overall positive and overall negative tweets and calculated the viewshed. Using this approach the output of the Observation Point Tool for each point is a binary raster dataset which is at the same spatial resolution of the input digital elevation data (90 m). This dataset signifies whether a particular cell is either visible (1) or not visible (0) from the point analysed. The percentage of visibility was calculated from the output raster based upon the number of raster cells classed as visible and the total number of cells within the >20 km radius of the point.

To facilitate the further viewshed analyses an input mask was derived from the viewshed results of each point. This involved using the Raster Calculator (ESRI, 2011) to convert the non-visible cells from ‘0’ to ‘NoData’ so that they could be ignored in further calculations. This viewshed input mask would be used to in further analyses to determine the land cover, ruggedness, and ESS agreements within the viewshed of each tweet.

### Land Cover within viewshed

The 2007 LCM (Morton *et al*., 2011) was also spatially clipped to within 25 km of the PWNT to accommodate the 20 km viewshed of each of the trail users’ tweets. The viewshed input masks and LCM (Morton *et al*., 2011) were combined using the Raster Calculator (ESRI, 2011) to determine the majority land cover class within each tweet’s viewshed. This was achieved by multiplying the viewshed input mask by the LCM (Morton *et al*., 2011). The result of this was the categorical land class(es) within the viewshed, and ‘NoData’ values for areas outside the viewshed. The majority statistic for each output was then calculated. This process was automated through use of ModelBuilder (ESRI, 2011) to iterate through each of the viewshed input mask datasets.

### Ruggedness within viewshed

Striking topographic features and challenging terrain are viewed as qualities of wild land (Carver, 2008). To quantify this a Ruggedness Index was devised. Ruggedness was derived from the SRTM DEM (Pope, 2009). Based upon a comparison of methods for calculating the ruggedness of landscape conducted by Cooley (Unkown), and work by Ascione *et al*. (2008), the standard deviation of elevation was chosen as a proxy for the ruggedness of the landscape.

The standard deviation of the DEM was calculated using the Focal Statistics tool within ArcMap (ESRI, 2011). This calculated the standard deviation of the DEM within a 3 x 3 cell raster window (equivalent to 270 m on the ground). This moving window passed over the 25 km PWNT DEM.

The resulting output was a raster dataset with float (decimal) values for each cell. This output was reclassified into equally divided quintiles, thus providing relative ruggedness across the 25 km PWNT corridor study area. The values where then standardised to a scale of between +1 and +5. This provided the Ruggedness Index.

The ruggedness for the viewshed of each overall positive and overall negative tweet was calculated through use of the Raster Calculator (ESRI, 2011). Firstly, the viewshed input masks and ruggedness dataset were used to calculate the Roughness Index within the viewsheds (areas outside the viewshed were assigned ‘NoData’). Secondly, the mean cell roughness of each viewshed was calculated to provide an average measure of ruggedness within the viewshed.

### ESS agreements within viewshed

To determine the extent of ESS agreements within each viewshed, the 25 km ESS agreement shapefile was first converted into a raster dataset. The cell size was set to 25 m, which was consistent with the LCM (Morton *et al*., 2011). A smaller cell size was chosen to limit the loss of detail caused when transitioning from a vector to raster format that may affect the boundaries between agreements.

In Raster Calculator, the ESS agreement raster was multiplied by the viewshed input mask. The extent of each ESS agreement type (ELS, HLS etc.) was then calculated as a percentage of the total viewshed for each of the overall positive and overall negative tweets.

# Results

* The sentiment expressed in the twitter messages along the PWNT trail corridor:
  + 47 tweets expressed positive sentiment,
  + 22 negative sentiment
  + 102 expressed no sentiment (i.e. neutral).
  + 10 tweets expressed both positive and negative sentiment.
  + Overall tweet sentiment consisted 40 positive, 16 negative and 105 neutral.
  + 105 tweets did not convey any sentiment and were classed as neutral. It was discovered that of these 105 tweets, 94 contained a URL within the TweetText.
    - Missing sentiment in images?
  + [Link to interactive map of trail user tweets](http://tom-wilson.info.s3-website-us-west-2.amazonaws.com/PennineTweets.html)

# Discussion

There are several issues with social web data that deserve attention. First is the accessibility to the data. To use Twitter as an example, none of the public APIs provide direct, unfettered access to Twitter data. The Twitter API is believed to be subject to a ‘streaming cap’ of about 1% of all tweets at any point in time (Driscoll and Walker, 2014). In reality it is only the social web companies themselves that have full access to the data (Manovich, 2011), and at the same time have full control as to who can access the data (boyd & Carwford, 2012). So of any data collected using the public Twitter API there also exists an additional ~99% of data in not accessible. This data cannot be accounted for because of a lack of transparency regarding the exact streaming cap, and the process of selecting which tweets are available via the API (boyd & Crawford, 2012). Nevertheless it is important that this is recognised within the research.

Second is the representation of social web data. Twitter usage requires access to the internet via a desktop computer or a mobile device. As such, Twitter usage is limited to internet users. Moreover Twitter usage is not evenly distributed among internet users (Driscoll and Walker, 2014). Specific to this research, not all trail users are necessarily internet users, and even those that are internet users may not wish to tweet when out hiking. Furthermore, a person needs both a smartphone and a data plan in order to send a tweet whilst out hiking on the PWNT. Consideration of this is needed to avoid the creation of a ‘digital divide’ whereby only trail users with smart phones and data plans are heard.

Third is the question of ethics. As boyd and Marwick (2011) succinctly put it; “there is a considerable difference between being in public and being public” (boyd & Crawford, 2012 p.673). Although Twitter data may be classed as public data, consideration should be given to the subjects of the study. Twitter users should understand that their data is public (unless they specify otherwise in their preferences), but there is the chance that they may not. Even if they do know their data is public they may not intend for their data or tweet to become public (Eckert *et al*., 2013). The Twitter dataset in this study did not contain Twitter usernames or personal information, and no real tweets have been published in this report. It may be necessary to disclose the purpose of data collection in the pursuit of a social web opinion-mining campaign.

# Reccomnedations

A process to select trail users’ tweets from a larger dataset of Twitter and extract the sentiment conveyed has been developed. The exploratory analysis of the data used in this research did not provide conclusive results with regard to the effectiveness of the ESS, but it did uncover interesting insights which deserve further attention.

As previously mentioned, an interesting avenue of future research is to determine the extent of image-sharing in trail users’ tweets. Furthermore, are tweeting trail users attempting to convey sentiment through these images? Borth *et al*. (2013) present research on visual sentiment ontology which could provide a foundation of future research in this area.

Based upon the findings of this research it is recommended that Natural England proactively initiate a social media strategy as a method of eliciting the sentiment of its trail users from their social web data. Natural England should select a hashtag with which it would like users to tag their tweets. This research has uncovered that trail users already utilise hashtags within their tweets. Assigning a hashtag specific to this campaign would facilitate the grouping and selection of tweets during data analysis. Furthermore, the hashtag can form the basis of a promotional and educational campaign.

A campaign would provide the opportunity to increase awareness of the use of social web data for representing views to the organisation. This is important from both an ethical and representative perspective: Trail users should be alerted to the fact that the sentiment they convey and comments they make are public. In terms of representation, a greater number of people need to be encouraged to participate in this scheme for it to be anywhere close to representative. However, alternative mechanisms will be additionally needed to ensure that trail user opinions are not subject to a digital divide whereby only those with a smart phone and a data plan are able to offer their opinion.

Initiation of a social media campaign is also likely to increase the amount of data available for analyses such as those presented in this report, and allow for the process to be refined further.

# Conclusions

# ADDITIONS:

* more about the collection of the twitter data
* statistics in results
* broader potential impact of study
* section regarding social media analysis in general
* policy reccomendations

# References

Compiled - will be added last.