Technological Forecasting & Social Change 95 (2015) 96–108

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

# Detecting tension in online communities with computational Twitter analysis

Pete Burnap [a,*], Omer F. Rana [a], Nick Avis [a], Matthew Williams [b], William Housley [b], Adam Edwards [b], Jeffrey Morgan [b], Luke Sloan [b]

[a] Cardiff School of Computer Science & Informatics, Cardiff University, Cardiff, UK
[b] Cardiff School of Social Sciences, Cardiff University, Cardiff, UK

ARTICLE INFO

*Article history:*
Received 12 October 2012
Received in revised form 18 April 2013
Accepted 19 April 2013
Available online 11 May 2013

*Keywords:*
Opinion mining
Sentiment analysis
Text mining
Social media analysis
Machine learning
Conversation analysis
Membership categorization analysis

ABSTRACT

The growing number of people using social media to communicate with others and document their personal opinion and action is creating a significant stream of data that provides the opportunity for social scientists to conduct online forms of research, providing an insight into online social formations. This paper investigates the possibility of forecasting spikes in social tension – defined by the UK police service as "any incident that would tend to show that the normal relationship between individuals or groups has seriously deteriorated" – through social media. A number of different computational methods were trialed to detect spikes in tension using a human coded sample of data collected from Twitter, relating to an accusation of racial abuse during a Premier League football match. Conversation analysis combined with syntactic and lexicon-based text mining rules; sentiment analysis; and machine learning methods was tested as a possible approach. Results indicate that a combination of conversation analysis methods and text mining outperforms a number of machine learning approaches and a sentiment analysis tool at classifying tension levels in individual tweets.



## 1. Introduction

Social networking technologies, such as those provided by online social networking sites such as Facebook (where users have an online social "friendship" relationship), and microblogging websites such as Twitter (where the online social relationship can be uni-directional or reciprocal), provide people with unprecedented interactive opportunities. These services also support the dissemination, discussion and often distortion of information within and between online and offline 'communities' at speeds hitherto not possible. The blurring of the boundary between offline and online discourse requires us to rethink the definition of human-to-human interaction. In this paper we study this new form of interaction by examining how events that occur offline have the potential to affect terrestrial societal cohesion and order, and how they are reflected in discourse conducted through online social media. We argue that offline events can trigger 'online tension' that can be measured via a series of indices or metrics. These metrics can then be used to monitor peaks in tension in social media, affording a form of 'anticipatory governance'.

Tension is defined by the Police Service in the United Kingdom as "any incident that would tend to show that the normal relationship between individuals or groups has seriously deteriorated and is likely to escalate to wider groups other than those involved" [1]. It is therefore important to study the possibility that collecting, analyzing and visualizing self reported information posted to online social networking sites could provide an insight into online tension through the mining and analysis of public opinion.

Opinion mining of social media data has received a lot of research attention in recent years (e.g. [2–5]), largely due to the increasing use of social networking technologies that enable citizens to self-report their opinions as frequently as they wish on a wide range of topics, and the availability of programmatic access to such data through application programming interfaces (APIs). The method of opinion

* Corresponding author. Tel.: +44 2920874000.
*E-mail address:* p.burnap@cs.cardiff.ac.uk (P. Burnap).

0040-1625/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.techfore.2013.04.013

mining generally requires the identification of: an entity to which the opinion is focused on (e.g. a person); attributes of the entity (e.g. the person's political perspective); views, attitudes or feelings towards the entity and its attributes (commonly defined as sentiment); an opinion holder; and a time at which the sentiment was expressed [6].

One of the key research challenges for textual opinion mining has been to classify sentiment expressed by the opinion holder, or a collection of opinion holders (e.g. comments posted in response to a news story), using a pre-defined set of classes (e.g. positive, negative or neutral). To achieve this, it is typical to try and identify features within the text that are useful for deciding which class it belongs to. These features may be positive or negative words (e.g. good, excellent, bad, and awful) or descriptive words in the context of the entity (e.g. this person is wrong about this = negative). Features may also be syntactic, such as verbs or nouns, or words that belong to particular lexicons (e.g. common swear words). In social media there are idiosyncratic features such as hashtags (the # symbol) and emoticons, which can also be used to inform classification decisions.

The Cardiff Online Social Media ObServatory (COSMOS) is a Web Observatory platform developed to support researchers interested in collecting, analyzing and visualizing publicly accessible digital data feeds [7,36-37]. One specific objective of the COSMOS platform is to develop computational tools to mine opinion from social media data and enable the detection of tension indicators in online communities, visualizing them such that spikes in tension can be detected. A spike exists where tension can be measured at a single point in time as significantly higher than it has been previously — an anomaly in a timeline.

Using the definition of tension as defined in [1] – "any incident that would tend to show that the normal relationship between individuals or groups has seriously deteriorated and is likely to escalate to wider groups other than those involved" – this study is specifically focused on developing an opinion mining application capable of classifying public posts, published through the social media microblogging site Twitter, with respect to the level of tension expressed. In terrestrial communities, such as particular neighborhoods and geographic regions, tension indicators are often obvious and apparent to the naked eye. Broken windows, graffiti, banners and posters are all potential indicators of tension within a locality. Tension indicators in online communities are not so obvious, but Twitter is a very open self-reporting platform that is underpinned by an online social network and has recently been shown to host discussions around sensitive topics such as riots [8], political campaigns [4], and the Tunisian and Egyptian revolutions [9,10]. Hence, messages posted to Twitter – 'tweets' – are the device through which it is proposed that tension can be detected.

The premise of the study from a forecasting perspective is that it will be possible to detect a rise in tension levels over time in relation to an observed event. Such a rise would thus be indicative of growing online tension, and allows those observing online tension to react accordingly. With the growing use of social media for expressing opinion, and the relative uncertainty around its interpretation, it is expected that a visualization of tension classified at a number of levels, with each level being plotted over time, would facilitate the observation process. From this, increases in the plotted tension levels – particularly the higher levels – are visible and therefore could be used as a metric for use in forecasting disruptive social phenomena during, and following, a known event.

In this paper we present the results of a tension classification study using a number of computational approaches. In Section 2 we provide a summary of the state-of-the-art in computational sentiment analysis and its application to study social phenomena. In Section 3 we discuss data collection and annotation along with the various methods used in the study and how they build on existing best practice. In Section 4 we discuss the results of applying different computational techniques to the problem of tension classification. In particular, the aim of this task was to test the performance of a number of content analysis and machine learning approaches with the intended outcome of identifying the best performing methods for classifying tension on a number of levels. We were motivated by the assumption that no single method would perform best at every level of tension. Finally, we discuss the benefits, relevance, and limitations of the outcomes in Section 5.

## 2. Background

Qualitative and quantitative research have been re-framed in the light of the rise of Web 2.0. Rathi and Given propose a framework for research in Web 2.0 environments (Research 2.0), which considers the web as a research platform, harnessing the power of crowds, and creating research databases from expansive online content that they refer to as perpetual data and coin the term Data 2.0 [11]. Data can be obtained from many sources such as personal blogs, wikis, microblogs posted on social networking sites (e.g. Facebook and MySpace status updates and Twitter tweets) and RSS (really simple syndication) feeds from sources such as news reporting websites [12].

These data, combined with emotive opinion mining methods, have been used in several recent sociological and economic academic studies. Computational statistical data mining methods have been used to investigate the demographics and 'friending' behavior on MySpace [13,14]. Tailored social media data harvesting methods, combined with human interpretation, have been used to investigate the change in people's ideology concerning the privacy of childbirth [15]. In [5] the authors conduct a large scale study of online social networks to extract different types of mood using sentiment analysis and the Profile of Mood States (POMS) psychological mood scale, and relate changes in mood to real-world socio-economical events. Building on this, psychological "well being" states have been monitored over time to show that online social networking reflects the assortive nature exhibited offline [2]. Sentiment analysis has been used to determine emotional differences between genders on MySpace [13,14], and study levels of positive and negative sentiments in Facebook [16] and Twitter comments [2,3]. Sentiment analysis has also been used to 'predict' election outcomes [4], and it was demonstrated that sentiment relating to new movie releases, combined with tweet frequency, was more accurate at predicting revenue than the Hollywood stock market [17]. For forecasting purposes sentiment analysis appears to offer a significant insight into online communication.

Three common approaches to sentiment analysis are text-based machine learning, lexicon-based methods and linguistic

analysis [3]. For machine learning, previous research has shown that probabilistic machine classifiers such as Naïve Bayes (NB) and support vector machines (SVMs) stand out as having produced the best results [18–20]. Probabilistic classifiers make classification decisions based on the learned probability that unseen text belongs to one of a number of pre-defined categories based on the features it contains. The classification decision is underpinned by Bayes Theorem, which formulates the probability of a new data item belonging to a particular class given previous knowledge of data items with a known classification. The weakness of this approach is that unseen data not learned during the training phase is not helpful in classification. Nevertheless, for short text such as tweets and sentences, NB has been shown to produce very good results [18–20]. Support vector machines have also performed well for short text. SVMs aim to classify unseen data by maximizing the distance between clusters of similar data points created using training data, and finding an optimal solution as to where to place new data points, and have been particularly useful for text classification [18,21,22].

The key to achieving success in text classification using machine learning is feature selection. Machine learning algorithms base classification decision on previous knowledge that certain features identify data as belonging to a particular class. For sentiment analysis, features typically include sets of single words, word pairs and word triples found in text [3]. Some research reports best performance for unigrams (single word features) [23], while other works report that bi-grams and trigrams (two and three word combinations) perform better [24]. Lexicons and linguistic analysis have also been used as machine learning features to try and improve results. For example, linguistic features such as parts-of-speech (e.g. noun, and verb) and lexicons of negation terms (e.g. not, and never) have been used for Twitter sentiment analysis [19]. These additional features did not improve results in this case but in other works they have resulted in better performance [25], where Twitter specific syntactic features such as hashtags (#), retweets (RTs) and user mentions (@) were also used as features. Other syntactic features such as emoticons have also been used to derive sentiment [18,20]. Lexicons of weighted terms, such as words that indicate a sentiment 'strength' have also been used to assist in text classification [26].

A recent HMIC (Her Majesty's Inspectorate of Constabulary) report that made 24 recommendations following G20 protests in April 2009 includes a recommendation that those within the police in charge of training, tactics, and community tension monitoring must be able to detect a situation in its infancy and make decisions on real time intelligence, such that they can react quicker while adapting to any inevitable change in a potentially socially disruptive situation. This involves much more pro-active intelligence gathering and monitoring [27]. Using generic sentiment analysis tools is an obvious solution to this problem. We are not aware of any study that has focused specifically on the classification of communication that could be indicative of a breakdown in societal cohesion. Therefore, in this work we aim to build on the existing text-based machine learning, and lexicon-based and linguistic analysis methods to develop algorithms for the identification of social tension and compare the resulting analysis with existing sentiment analysis tools.

## 3. Methods

Fig. 1 illustrates the overall method from data selection to result comparison. This section explains each part of the method in detail.

### 3.1. Data collection, selection and annotation

In this study the microblogging site Twitter was used as the source of data. Twitter provides tiered access to its data and is one of the most open forms of social media communication given its non-reciprocal friendship structure. This allows one-to-many communication, making it the nearest social media equivalent we have to the agora and public square. Other social media sites, such as Facebook, are very closed and access to public online discourse is more restricted.

As shown by the recent study of the UK riots [8] and Arab Spring [9,10], tension would be expected to occur in tweets during the period surrounding a particular event, such as a riot or political campaign. It was therefore decided to take tweets posted around a particular event as a study dataset. At the time of the research project there was a particularly pertinent event surrounding an allegation made by Manchester United footballer Patrice Evra that Liverpool footballer Luis Suarez racially abused him during a match between the two teams. Given the media frenzy surrounding this it was decided that a study focusing on racial tension would be appropriate, with the Suarez–Evra incident being the event around which tension could rise.

Twitter offers two application programming interfaces (APIs) for collecting tweets: one is the search API, which may be used to retrieve past tweets matching a user specified criteria; the other is the streaming API, which may be used to subscribe to a continuing live stream of new tweets matching a user defined criteria and delivered to the user as soon as they become available [28]. Researchers need not define any specific criteria to receive data from the streaming API. They can receive a (free) 1% sample of everything posted each day, as it is posted. The Cardiff Online Social Media ObServatory (COSMOS) collects data in this way, harvesting a sample of online discourse through Twitter that amounts to $>3$ million tweets per day. COSMOS enables real-time analysis as well as retrospective longitudinal study of discourse for events that were not anticipated and for which data was not collected at the time of the event. The search API only allows a user to retrieve data up to 14 days prior.

Data was extracted from the COSMOS archive for one month before and one month after the event. The reason for collecting data during the lead-in period and following the event was to identify the difference in tension-related sentiment before, during and after the event so that, in the future, we may be able to forecast raised levels of tension based on a rise above 'normal' levels. In opinion mining terms, 'Suarez' is the entity to which the opinion study is focused on and was therefore the key sampling term. As the event actually includes two actors, we could have extended the dataset to also include 'Evra', but as there may be cases where only one actor is present in the event, the tension detection engine was designed to require only one actor to be defined by the observer. To include other actors in an

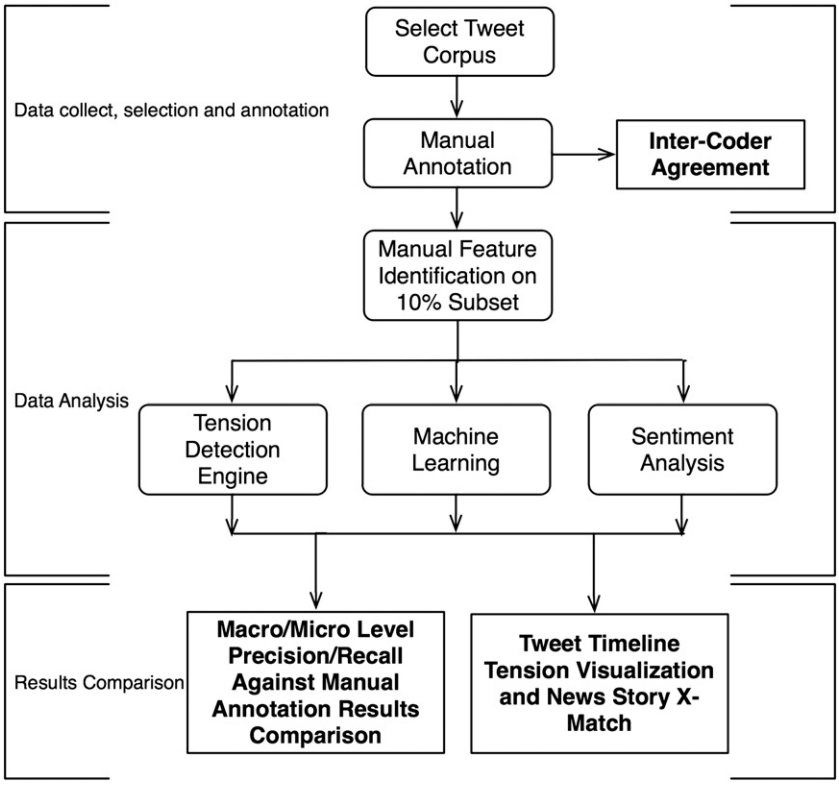


**Fig. 1.** Method overview.

analysis, we could run parallel tension detection processes (e.g. with 'Evra' as the actor). Any tweet posted during the specified period that included the term 'Suarez' was extracted from the archive. This would include any tweet using the Twitter topic classification device, often referred to as a 'hashtag' (e.g. #suarez, and #suarezisaracist). The sample produced a study corpus of 1022 tweets. This is a relatively small corpus for a machine learning classification task but in order to validate the results it was essential to derive a gold standard dataset that was coded according to the level of tension it contained. Human coders were used to provide some "ground truthing" for the data and it is common to use a smaller dataset where this is required [18,21,25].

Furthermore, it is useful to note that human coders themselves can also differ in their assessment of sentiment. The subject specific expertise and knowledge of the human coders are therefore essential to provide more objective assessment of the data. In this experiment the expert human coders were specialist police officers who focus on identifying tension in terrestrial communities every day. The study corpus was manually classified by four police officers using an ordinal scale from 0 to 3 where: 0 meant the tweet was irrelevant; 1 meant the tweet contained "some tension"; 2 meant the tweet contained "tension"; and 3 meant the tweet contained "high tension". The decision on the degree of tension the tweets contained was left to the coders' discretion, which meant there was a large degree of human subjectivity involved in classifying the tweets. Because of the subjective nature of the data classification exercise it was important to determine the degree to which the individual coders agreed, that is, how often they all coded a single tweet with the same level of tension. To calculate the level of inter-coder agreement we used Krippendorf's alpha coefficient [29] — a well known statistical measure of agreement between independent coders where a measure of 0 indicates no agreement and a measure of 1 indicates perfect agreement. Despite the flexibility of having three levels of tension to choose from, the inter-rater agreement calculated using Krippendorf's alpha was 0.67, which is deemed an acceptable level of agreement for drawing tentative conclusions from human annotated data [29]. In related sentiment analysis research, the agreement was around 0.57 [21]. We do not assume here that the humans are absolutely correct, only that they provide a gold standard for tool accuracy validation. The fact that they are trained in detecting tension through language use is a benefit for the validity of the classification exercise.

The 'irrelevant' class was added as we expected significant "noise" in the dataset. For instance, tweets coded as 'irrelevant' included those not written in English (we only focus on English text in this research), those containing the term "Suarez" but not related to the footballer, and others containing the term but relating to his performance/ability as a footballer — not relating to the event under study. This does not preclude 'irrelevant' tweets from containing tension, but for the purposes of tension detection experiments surrounding events we only wanted tweets relating to the event. We could have pre-processed the data to remove the "noise" but the ultimate aim is to deploy the tension classification tool to consume the real-time Twitter stream and produce real-time results, where pre-processing will not be possible.

The majority of tweets were not related to the event, with 52% of the 1022 coded tweets being coded as irrelevant by our coders. A further 306 (30%) were coded as having "some tension", 132 (13%) as "tension", and 48 (5%) as "high tension". We therefore had a significant number of tweets in the "tension" and "high tension" classes that had the potential to be distinguished from the rest and indicate spikes.

### 3.2. Data format and analysis

It is useful to identify the structure of a 'tweet' to better explain the subsequent data analysis carried out on this message. For a summary of the structure of a tweet see [28]. The authors point out that of these data text from_user (the text of the tweet) and time (the time the message was posted) are often the key components used in analysis. Indeed, in this research we focused only on the qualitative text content and time the tweets were sent. This allows us to identify the entity and opinion expressed towards it at any given time. We could also identify the opinion holder through additional tweet attributes if required.

The problem of identifying various levels of tension in text is essentially a classification task. Having had the corpus of tweets manually coded, the objectives of the task were to develop software that can (i) take each tweet and correctly classify it by coding it at the same point on the ordinal scale as the gold standard produced by manual coders and (ii) plot a longitudinal visualization of tweets over time to determine if spikes in tension were visible — firstly in the human annotated data, and subsequently by using various computational methods. The computational methods underpinning each type of analysis are discussed in this section.

#### 3.2.1. Test set feature identification

A random 10% sample of the coded data was manually inspected, with the aim of identifying features that might distinguish tweets containing different levels of tension.

It was clear from such a small sample that tweets manually coded as containing 'some tension' were mainly used to disseminate information within the online community. Very few of these tweets actually expressed any opinion and many of them contained a hyperlink as a reference to support the information the user was sharing. Another obvious feature was that tweets coded as containing 'high tension' included what one may call extreme expletives — English words 'F***' and 'C***'. 'High tension' tweets also appeared to contain expletives (of a less extreme nature) used together with accusations related to the event (e.g. "…is a racist t**t", and "…is a total bloody bigot…").

Tweets coded as 'tension' had less obvious features from a small sample but clearly expressed opinion as opposed to simply passing on information. They exhibited expletives, derogatory (racist) terms and accusations related to the event, thus where either of these were used mutually exclusive in a tweet, an assumption was made that the tweets would contain 'tension', unless the tweets contained a hyperlink, in which case it was labeled as 'some tension' based on the earlier observation.

Even in such a small test set, a taxonomy emerged from the inspection of tweets that is related to the typology of the tweet. It is clear that there are three types of tweet in relation to tension surrounding an event — passive-informative (where information is simply passed on), opinionated (where some opinion was given but not particularly forceful), and aggressive (where extreme opinion and anger were present).

#### 3.2.2. Tension analysis engine

To understand the development of a tension classification tool, let us start with a working definition of opinion as a quintuple (e, a, s, h, and t), as defined in [6], where e is an entity representing the target of an opinion (e.g. a person); a is an aspect of e (e.g. the person's political perspective) — note that each e can have more than one a; s is the sentiment towards a, where s can be positive, negative or neutral and be expressed with different levels of strength (e.g. −5 (−ve) to +5 (+ve)); h is the holder of the opinion; and t is the time at which the opinion was expressed. This framework for opinion mining is typically used for sentiment analysis.

For the purposes of analyzing tension expressed through the medium of social media, each tweet can be framed as an opinion where e is the entity representing the target or source of the tension (e.g. a person such as Suarez); a is an aspect of the entity (e.g. an action performed by e that has led to the tweet response — such as an action of alleged racial abuse); s is a level of tension in relation to e, a or both, ranging from 0 (irrelevant) to 3 (high tension); h is the Twitter user that posted the tweet; and t is the time it was posted. By plotting the value of s for n tweets over a period of time ranging from t(min) to t(max), we aimed to identify points in time where s is significantly higher at t than it was previously. We can define these points as spikes in tension.

The first challenge was to handle the "noise" in the dataset. Twitter is a highly discursive platform and it is likely that tweets could be referring to other instances of the entity. For instance, there are many other people called Suarez using Twitter who could be mentioned in a tweet. Entities can also have many aspects. For instance, the aspects of Suarez also include his ability as a footballer, his position as a national team player etc. Thus we must ensure that the tweets are referring to the correct instance of entity e, and in the correct context — the context defined by the chosen aspect of the entity a. This is how we distinguish tweets belonging to the irrelevant class from those containing tension. To achieve this, we return to the earlier observations of the random sample where it was evident that tweeters (users posting tweets) were connecting actors to events (e.g. "…Suarez charged for racism"), assigning attributions (e.g. "Suarez is a racist"), and stating accusations (e.g. "…Suarez abused him…", and "…Suarez called him a…"), within their naturally occurring conversation. This logically led us to the conversation analysis methods of Harvey Sacks, in particular, membership categorization analysis (MCA) [30].

Sacks's work concerned the close empirical analysis of everyday naturally occurring talk-in-interaction. Thus, Sacks' 'analytic' mentality displayed a concern with the plethora of description that everyday language exhibits. Furthermore, descriptions occur within a wide range of discursive contexts. For example, newspapers, business meetings and school lessons all provide for the generation of descriptions albeit within different contextual arrangements. Naturally occurring social media updates e.g. 'tweet formulations' are no different in this respect.

For Sacks, one of the important features of conversation and description is the display of categories and the methodical process of categorization. In Sacks' famous example 'the baby cried the mommy picked it up' these considerations are illuminated by an analytical consideration of how we make sense of the story. In terms of Sacks' example we understand the story in terms of the 'mommy' picking up her 'baby' in response to the baby 'crying'. For Sacks, we understand the story in this way because we associate the categories of 'baby' and 'mommy' with the membership categorization device (MCD) 'the family'. Sacks used the story, along with the pre-described rules of application (the economy and consistency rules), to generate a set of analytical concepts, namely membership categorization devices (MCDs), membership categories (MCs) and category bound activities or attributions (CBAs). Personal categories such as 'mother', 'father', 'son' or 'daughter' are described by Sacks as MCs. Furthermore, they are viewed as membership categories of the MCD 'family'. In addition to this framework the category machinery was complemented by the notion of CBAs. According to Sacks, they can be understood as an attempt to describe how certain activities or attributions are commonsensically tied to specific categories and devices (e.g. in the case of Sacks' story the tying of the activity of crying to the category 'baby'). The theory here is that for the purposes of identifying an association between an entity mentioned in a tweet (e) and a particular aspect (a), we can use membership categorization analysis as a useful term linkage framework by considering an entity (e) as an MC, and deriving a lexicon of action, attribution and association terms related to an aspect (a) of the entity. When an MC and a CBA term co-occur in a tweet, they enable us to link the entity to the chosen aspect of the entity, and thus to the event that prompted the opinionated response, with the event becoming an MCD. In this sense MCA is being used to inform socio-technical design [34].

Given the organizational characteristics of social media updates (e.g. the 140 character length tweet) membership categorization analysis provides an empirically tested approach to the understanding of natural language practice, which relies on social explication and inter-subjective interaction as opposed to occluded internal processes. It is a model of human category and associated linguistic work based on the systematic observation of everyday human practice and empirical materials [31–33]. As a consequence it provides a means of developing some bespoke rules that can inform the analysis of large scale social media data that can provide a way of analyzing content in terms of category configuration and different forms of attribution and activity.

With the classification task in mind, experiments were conducted to determine if the principles of MCA were applicable to identifying whether tweets were specifically related to an event, and whether mentions of the entity were contextualized within a given aspect. We assumed the MCD to be the event itself (i.e. the Suarez–Evra incident) and that for a tweet to be relevant to the event it should contain at least one reference to an MC associated that MCD (i.e. "Suarez", "Evra"). All tweets contained "Suarez" as he is the entity against which we are mining opinion and hence it was the keyword used to select the data from the COSMOS Twitter archive, but the presence of "Evra" confirms the MCD presence through the notion of standardized relational pairs [33] — the presence of this pair indicates an association of the tweet to an event, in the same way 'mommy' and 'baby' are associated the MCD 'the family'. Where only one MC was mentioned, we included the MCA concept of CBAs to further examine the tweet. CBAs include predicates that are used in relation to an MC. Based on the observations from the sample of the coded dataset, it was evident that CBAs relate very well to the accusation terms used in tweets (e.g. "… Suarez abused him…", and "…Suarez called him a…"), or attribution claims that one of the MCs (Suarez) conducted such actions and therefore "Suarez is a racist". CBAs within tweets were used to classify tension by weighting different predicates used in association with an MC, with different types of CBA being given a different weighting. In the first instance, mentions of an MC and an activity or attribution predicate, such as "racism", "racist", and "called" (for full CBA list of terms see algorithm in Fig. 2) were weighted as being relevant to the event (i.e. not in the 'irrelevant' class). Tweets with a single MC but without a CBA feature were classified as 'irrelevant'.

The next stage was to distinguish between levels of tension. At this stage additional features and lexicons relating to the 'strength' of opinion were used to identify levels of tension in a tweet. If a hyperlink was present in the tweet (e.g. "Suarez has been charged http://…"), it was used as evidence to suggest that the tweet was more likely to include passive information (i.e. 'some tension') than be of a forceful individual opinion. Tweets featuring an MC, a CBA, and a mild expletive (using an email swear word filter list as a lexicon of expletives) were classified as 'high tension' based on observations from the sample of coded data (e.g. "Suarez is a racist t**t"). A small number of 'extreme' expletives were also defined, the presence of which automatically led to the classification of a tweet as 'high tension' if the previous MC + MC (standardized relation pair) rule was true. The combination of expletive predicates used in conjunction with a CBA, or the sole of an 'extreme' expletive and more than one MC, were therefore weighted as 'high tension'. Where a racist predicate (using an online racist term database as a lexicon for racist predicates) ("Evra is a black…") was present in a tweet it was classified as 'tension'. If none of the additional features were present, the tweet was marked as 'tension' based on the presence of an MC and a CBA being indicative of attributions and associations being assigned to the entity (e.g. "…Suarez is a racist"…).

Based on these features an algorithm was defined for the codification of the MCA principles, which is detailed in Fig. 2. This algorithm effectively represents the method used to classify tension with a tension classification engine. This implementation is specific to racism but the advantage of using the conceptual framework of MCA allows future developments to the tension classification engine to include other types of tension (e.g. homophobic, economic, and political) by changing the MC and CBA lexicons to include different entities and attributions, and using different sentiment 'strength' lexicons, such as derogatory terms used against a particular type of entity. It could be possible that the algorithm could be modified to classify and detect tension around other events such as politics (e.g. where the MC list contains political candidates and the CBA includes "elitist", "corrupt", "lied", etc.) or policing (e.g. where the MC list and derogatory term lexicon

Tweets (T) = corpus of tweets
Membership Categories (MC) = {suarez, evra}
Category Bound Activities or Attributions (CBA) = {racism, racist, racial, called, calling, guilty, innocent, punish, discriminat*, xeno*}
Expletives E = {lexicon of expletives taken from an online swear word list[2]}
Extreme Expletives EE = {two expletives beginning with 'f' and 'c'}
Racist terms R = {lexicon of racist terms taken from the online racial slur database[3]}

For each tweet *t* in *T*

1. Count the number of instances in *t* of membership categories *mc* from *MC*
2. If *mc* > 1 → we are confident *t* is related to the event (go to 4)
3. If *mc* = 1 → we need to perform an additional check to ensure *t* is related to the event
    a. Check if *t* contains a Category Bound Activity or Attribution term *cba* in *CBA*
    b. If *cba* >= 1 we are confident this tweet is related to the event (go to 5)
    c. If *cba* <1 we are NOT confident *t* is related to the event and mark it as IRRELEVANT
4. Check if *t* contains an extreme expletive *ee* from *EE*
    a. If *ee* >= 1 mark *t* as HIGH TENSION and finish
5. Check if *t* contains a hyperlink
    a. If a hyperlink is present mark *t* as SOME TENSION and finish
6. Check if *t* contains an expletive *e* from *E*
    a. If *e* >= 1 mark *t* as HIGH TENSION and finish
7. Check if *t* contains a racist term *r* from *R*
    a. If *r* >=1 mark *t* as TENSION and finish
8. Else mark *t* as TENSION

**Fig. 2.** Tension detection algorithm.

contain various names used to describe police and CBAs include terms surrounding tactics and the handling of crime, etc.).

### 3.2.3. Machine learning

In addition to the codification of the tension analysis engine, a machine learning approach to classifying the corpus of tweets was also implemented. As discussed in Section 2, previous research into sentiment analysis using Twitter as a source of data has indicated that Naïve Bayes (NB) and support vector machine (SVM) approaches have performed best [18–20]. In particular, n-gram approaches using one, two and three word combinations as features have been successful with probabilistic methods [23]. n-Grams can be considered in terms of their frequency of occurrence, or simply in terms of their presence. The NB machine classifier is a suitable machine-learning algorithm for this approach because it supports multiple n-gram combinations and considers both presence and frequency of n-grams. Therefore a multinomial Naïve Bayes classifier was used to classify the dataset, supported by the WEKA machine learning toolkit[1], with unigram (single word), bigram (two words) and trigram (three words), and combined (uni-, bi- and tri-gram) features, with separate experiments for n-gram presence and frequency.

As well as n-grams, the presence and frequency of the lexical and syntactic MCA features as identified and implemented in the tension analysis engine (i.e. CBAs, expletives, racist terms and URLs) were also used to train separate machine classifiers (one experiment for presence, another for frequency). The sequential minimal optimization (SMO) algorithm for training a support vector classifier was implemented, as was a linear logistic regression (LLR) classifier. Logistic regression is used less frequently since SVMs were developed because SVMs are generally accepted to be more accurate as the approach is less concerned about 'fitting' the data and more tailored to maximizing accuracy. However, we were interested to determine if a linear method would be appropriate for tension analysis, so both were tested. A ten-fold cross validation approach was used to train and test the machine learning methods. This approach has previously been used for building machine classifiers for short text (e.g. [21]). It functions by iteratively training the classifier using 10% of the manually coded dataset, and classifying the remaining 90% as 'unseen' data, based on the features evident in the cases it has encountered in the training data. It then determines the accuracy of the classification process and moves on to the next iteration, finally calculating the overall accuracy.

### 3.2.4. Sentiment analysis

When studying the emotive content of tweets it is important to consider the performance of other widely used opinion mining tools. As a comparator to the tension analysis engine we used SentiStrength, a sentiment analysis tool that has been tested and evaluated using short text from social media [21] and is available for academic research. SentiStrength was not developed to distinguish any sort of relevance in relation to an event, so the sentiment analysis experiment was only conducted using tweets manually coded as containing some degree of tension (i.e. not those coded as 'irrelevant'). SentiStrength classifies text on a scale of $-5$ to $+5$ depending on the degree of negative to positive sentiments. We assumed extremes of positive or negative sentiments would indicate tension, and a number of mapping variations were trialed to relate the $\pm$0–5 measure of SentiStrength to the ordinal scale of the tension analysis engine. SentiStrength's best performing mapping was when $\pm 4$ or 5 was mapped to 'high tension', $\pm 3$ was mapped to 'tension' and $\pm 1$ or 2 was mapped to 'some tension'. 0 was mapped to 'some tension' and 'irrelevant' but reduced the performance in both cases. Mapping only the higher negative sentiment scores (i.e. $-4/-5$) to 'high tension' and 'tension' was also trialed but with less accuracy than mapping both positive and negative extremes to the higher levels of tension.

[1] http://www.cs.waikato.ac.nz/ml/weka/.

Furthermore, based on the assumption that extremes of positive or negative sentiments would indicate tension, we were interested in testing a hypothesis that for a set of tweets on a given day, the differential between the mean maximum positive and mean maximum negative sentiment scores on that day could also be used a tension indicator. That is, the bigger the difference between the two maximum scores (−5 to +5), the more tension exists, based on extreme opposite sentiment.

## 4. Results

The results of the classification tasks are provided using standard text classification measures of: precision (i.e. for class x, how often are tweets classified as x when they should not be — a measure of false positives); recall (i.e. for class x, how often are tweets not classified as x when they should be — a measure of false negatives); F-measure, a harmonized mean of precision and recall; and accuracy, the total correctly classified tweets normalized by the total number of tweets. The results for each measure range between 0 (worst) and 1 (best). Because of the specific interest in detecting spikes in tension there is particular interest in the accurate identification of 'tension' and 'high tension' tweets. Thus macroaveraged results for each class and microaveraged results for performance across all classes are presented. The tension analysis engine results are presented first (Section 4.1), followed by the machine learning approach (Section 4.2), and finally sentiment analysis (Section 4.3). The formulae for calculating these results are as follows (where TP = true positives, FP = false positives, TN = true negative and FN = false negative):

$$\begin{aligned} \text{Precision} &= TP/TP + FP \\ \text{Recall} &= TP/TP + FN \\ F - \text{measure} &= 2 \times ((P \times R)/(P + R)) \\ \text{Accuracy} &= (TP + TN)/(TP + FP + FN + TN) \end{aligned}$$

Following the discussion of the results, Section 4.3 presents visualizations of the dataset (i) as coded by police officers, (ii) as classified by the tension analysis engine, and (iii) as a function of the mean daily minimum and maximum daily sentiment. The purposes of the last section are (a) to illustrate the visual difference between (i) and (ii), i.e. how close does the tension analysis engine plot tension in relation to the "ground truth" data, (b) to visualize the similarity between spikes in tension and a large difference in mean maximum and minimum daily sentiments — to determine if differences in opinion using sentiment analysis are indicative of tension, and (c) to analyze media stories published on days when tension spikes to determine the possible "offline" triggers.

### 4.1. Tension analysis engine

Tables 1 and 2 contain the performance results of the tension analysis engine. At a microaveraged level the performance of the tension analysis engine achieved a precision and recall of 0.74 and an overall accuracy of 0.87, which are both encouraging scores for this type of classification. To better understand the ability of the tension analysis engine to detect high levels of tension, it is important to understand the performance at a macroaveraged level for each class.

Looking at the F-measure performance for the tension engine there is clearly room for improvement. There are a significant number of false positive and false negative classifications. 'Tension' is the weakest class at 0.51. This could be due to the fact that there were no "standout" features for the 'tension' class, as there were for the others. However, for an ordinal scale this is by no means a poor performance. Related work that has classified sentiment using Twitter data with a nominal positive, negative, and neutral scale achieved a similar result (around 0.5) [18]. The lack of distinct nominal classes arguably makes classification harder as there is more of a difference between positive and negative than between 'tension' and 'high tension'.

**Table 1**
Macroaveraged precision and recall results for the ordinal tension scale tension classification engine.

"Irrelevant"

| Machine | | Human Coders | |
|---|---|---|---|
| | | Yes | No |
| | Yes | 501 | 129 |
| | No | 35 | 357 |

Precision = 0.80
Recall = 0.93
F-Measure = 0.86
Accuracy = 0.84

"Some Tension"

| Machine | | Human Coders | |
|---|---|---|---|
| | | Yes | No |
| | Yes | 163 | 68 |
| | No | 143 | 648 |

Precision = 0.71
Recall = 0.53
F-Measure = 0.61
Accuracy = 0.79

"Tension"

| Machine | | Human Coders | |
|---|---|---|---|
| | | Yes | No |
| | Yes | 65 | 58 |
| | No | 67 | 830 |

Precision = 0.53
Recall = 0.49
F-Measure = 0.51
Accuracy = 0.88

"High Tension"

| Machine | | Human Coders | |
|---|---|---|---|
| | | Yes | No |
| | Yes | 26 | 12 |
| | No | 22 | 962 |

Precision = 0.68
Recall = 0.54
F-Measure = 0.60
Accuracy = 0.97

**Table 2**
Microaveraged precision and recall results for the ordinal tension scale tension classification engine.

| "All Codes" | | | |
|---|---|---|---|
| | | Human Coders | |
| Machine | | Yes | No |
| | Yes | 755 | 267 |
| | No | 267 | 2797 |
| Precision = 0.74<br>Recall = 0.74<br>F-Measure = 0.75<br>Accuracy = 0.87 | | | |

For sentiment analysis, most existing researches tend to use the accuracy score. This represents the measure of computational classifier and human annotator agreement. The macro-averaged results indicate that the tension analysis engine has an accuracy of 0.97 for 'high tension' and 0.88 for 'tension'. This is very encouraging for the purposes of classifying 'tense' tweets that would contribute to detecting a spike. It demonstrates that the tension engine agrees with humans almost all the time for 'high tension' tweets. It must be noted that for datasets with relatively small subsets, such as the 'tension' (13% of the overall dataset) and 'high tension' (5% of the overall dataset), a default classification choice to assign a new data item to any class other than the smaller classes is still likely to be relatively accurate. For instance, if the tension classification engine classified every tweet as 'irrelevant', it would still be around 50% accurate. However, this is generally an issue for probabilistic classifiers and not rule-based approaches such as the tension analysis engine where there is no 'most probable' option.

### 4.2. Machine learning

The Naïve Bayes classifier was tested using multiple experiments using unigrams, bigrams and trigrams. Separate experiments were performed for each n-gram length using n-gram presence and frequency as features. Using unigrams and feature presence produced the best performance (thus only results from this experiment are presented in the comparison to other methods), followed by the combination of uni-, bi- and tri-gram with term frequency ($P = 0.690$, $R = 0.722$, and $F = 0.683$), then bigram with term frequency ($P = 0.632$, $R = 0.673$, and $F = 0.647$), and trigram presence ($P = 0.610$, $R = 0.649$, and $F = 0.612$). Naïve Bayes performed best overall out of the three machine learning classifiers, suggesting that the words in the tweets are actually more accurate than MCA features for machine learning. The results for this method are macroaveraged in Table 3 and microaveraged in Table 4. The macroaveraged results for all three methods in comparison to the tension analysis engine are in Table 5. While it may be typical to average across all computational methods to achieve the best overall result, we do not do so in this case because of the focus on the performance of each classifier on an individual basis due to their different approach (rule-based, probabilistic and statistical best-fit).

While the results of the Naïve Bayes approach are very similar to the tension analysis engine at a microaveraged level, it is clear that at the macro level the ML approach is much less effective at detecting the raised levels of tension (i.e. 'tension' and 'high tension'). The Naïve Bayes method failed to classify any tweets as 'high tension'. As Table 5 shows, the SVM and logistic regression methods also performed very badly overall at the higher levels of tension.

The 'tension' and 'high tension' levels of tension are much less frequently occurring in the tweet corpus (13% and 5% respectively as a proportion of the total dataset), which would clearly have an impact on classification performance for supervised learning methods (i.e. those that 'learn' from the data), and even more so for probabilistic approaches. For instance, in probabilistic terms, it is highly improbable that a new tweet belongs to the 'high tension' class. It is accepted that the dataset is relatively small for machine learning tasks but we also note that it is not uncommon to use manually coded text datasets of around 1000 data points for testing

**Table 3**
Macroaveraged precision and recall results for the ordinal tension scale Naïve Bayes Multinomial Text Machine Classifier.

| "Irrelevant" | | | |
|---|---|---|---|
| | | Human Coders | |
| Machine | | Yes | No |
| | Yes | 490 | 101 |
| | No | 46 | 385 |
| Precision = 0.83<br>Recall = 0.91<br>F-Measure = 0.87 | | | |

| "Some Tension" | | | |
|---|---|---|---|
| | | Human Coders | |
| Machine | | Yes | No |
| | Yes | 243 | 147 |
| | No | 63 | 569 |
| Precision = 0.62<br>Recall = 0.79<br>F-Measure = 0.69 | | | |

| "Tension" | | | |
|---|---|---|---|
| | | Human Coders | |
| Machine | | Yes | No |
| | Yes | 36 | 12 |
| | No | 96 | 878 |
| Precision = 0.75<br>Recall = 0.28<br>F-Measure = 0.41 | | | |

| "High Tension" | | | |
|---|---|---|---|
| | | Human Coders | |
| Machine | | Yes | No |
| | Yes | 0 | 1 |
| | No | 48 | 973 |
| Precision = 0<br>Recall = 0<br>F-Measure = 0 | | | |

**Table 4**
Microaveraged precision and recall results for the ordinal tension scale tension Naïve Bayes Multinomial Text Machine Classifier.

| "All Codes" | | | |
|---|---|---|---|
| | | Human Coders | |
| Machine | | Yes | No |
| | Yes | 769 | 261 |
| | No | 253 | 2805 |
| Precision = 0.75<br>Recall = 0.75<br>F-Measure = 0.75<br>Accuracy = 0.87 | | | |

text analysis engines [18,21,25]. In future work we may oversample the instances of the higher tension levels to determine if increasing the ratio of high tension tweets increases the performance of the machine learning algorithms. There were instances where the machine performed better than the tension engine, for instance the precision of Naïve Bayes for the 'tension' class and SVM for the 'high tension' class is actually better, suggesting that less false positives are being produced with machine learning. However, the recall is much less, meaning many more tweets that are required to indicate a spike in tension are not picked up. The mixture of results here suggests that, in the future, a combination of these methods may improve tension detection overall where the method that performed best at each level of tension is used as the classifier for that class.

### 4.3. Sentiment analysis

Tables 6 and 7 provide comparative performance results for the SentiStrength sentiment analysis tool and the tension analysis engine respectively. Note that there is no 'irrelevant' category for this experiment, as SentiStrength is not designed to detect relevance. The F-measure is similar for both tools for the 'some tension' class, but the tension analysis engine performs better at the higher end of the tension scale, which makes it more useful for detecting raised levels of tension. These results suggest that extremes in positive and negative sentiments are not directly related to tension and that tension detection requires more than sentiment analysis alone.

### 4.4. Visualizing the data

The final experiment was to visualize the longitudinal levels of tension in relation to an event, with the aim of plotting the frequency of tweets classified as belonging to one of the four levels of tension over time. For forecasting purposes, it is proposed that the plots would enable people observing online tension in relation to an event to identify "spikes" in tension (i.e. high frequency) — particularly the high levels of tension. The baseline for forecasting purposes is therefore the "normal" frequency of tweets classified at each level of tension — starting from the time the event was first observed. Deviations from this baseline (i.e. a rise in the plot) could be used as a source of information to forecast raised tension. We note that, in this research, tension is directly related to an event. The tension detection engine does not yet determine what constitutes a "tense" event, rather it must be prompted to analyze tweets for a specific named entity (e.g. Suarez), and the outputs would be visualized as follows in this section. Tension is therefore a product of an event and is not forecasted prior to an event, rather within or following the event. As we will show, tension exists at various points in time following the event — but not persistently. There will be points in time where tension 'spikes', and there is a level of uncertainty as to when this may occur. The tension detection engine offers a level of intelligence as to when spikes may occur.

The plots are visualized as follows: (i) as annotated by the manual coders, (ii) as classified by the tension analysis engine, and (iii) as a daily differential between the maximum positive and negative sentiment scores. Image 1 shows the number of tweets manually classified in each class of tension per day. Spikes in tension can be defined as an increase in the number of tweets coded as 'some tension', 'tension' and 'high tension' in relation to the previous day (or days). By this definition, spikes can be seen on a number of occasions — most clearly on the 20th November 2011, 20th/21st December 2011, and 27th January 2012.

Cross-referencing the visible spikes with news stories on those dates (achieved by searching Google news aggregator with the term 'Suarez') reveals that Suarez was initially charged with racial abuse on the evening of the 16th November. There is a small fluctuation in "high tension" tweets on this day and a large spike in 'some tension' tweets in the following few days. This is logical given that 'some tension' tweets generally follow a passive information transfer construction with people passing on news stories and links as stories unfold. Suarez was found guilty of racial abuse and subsequently banned for eight matches on the evening of December 20th 2011. This correlates precisely with a large spike in tweets of all tension levels and is the highest peak on the 'tension' and 'high tension' lines. Note that the spike in 'high tension' follows the conviction and not the accusation, suggesting that evidence is required in online communities before tension spikes. This relates to the recent Guardian study investigating rumor spread on Twitter [8]. Once confirmation of rumors exists, this could have an impact on rising tensions.

**Table 5**
Macroaveraged performance results for the COSMOS tension engine and three machine learning methods (bold = best performance).

| | Irrelevant | | | Some | | | Tension | | | High tension | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | R | F | P | R | F | P | R | F | P | R | F |
| COSMOS | 0.80 | 0.93 | 0.86 | **0.73** | 0.58 | 0.61 | 0.53 | **0.49** | **0.51** | 0.68 | **0.64** | **0.60** |
| MNNB | **0.83** | 0.91 | **0.87** | 0.62 | **0.79** | **0.69** | **0.75** | 0.28 | 0.40 | 0 | 0 | 0 |
| SVM | 0.58 | **0.99** | 0.73 | 0.46 | 0.07 | 0.11 | 0.43 | 0.17 | 0.25 | **0.71** | 0.10 | 0.18 |
| LLR | 0.61 | 0.98 | 0.75 | 0.38 | 0.07 | 0.12 | 0.4 | 0.21 | 0.29 | 0.48 | 0.21 | 0.29 |

**Table 6**
SentiStrength performance against human coded "tension" indicators.

| "Some Tension" | | Human Coders | |
|---|---|---|---|
| Machine | | Yes | No |
| | Yes | 184 | 94 |
| | No | 122 | 86 |

Precision = 0.66
Recall = 0.60
F-Measure = 0.63
Accuracy = 0.56

| "Tension" | | Human Coders | |
|---|---|---|---|
| Machine | | Yes | No |
| | Yes | 31 | 61 |
| | No | 101 | 293 |

Precision = 0.34
Recall = 0.23
F-Measure = 0.28
Accuracy = 0.67

| "High Tension" | | Human Coders | |
|---|---|---|---|
| Machine | | Yes | No |
| | Yes | 19 | 97 |
| | No | 29 | 341 |

Precision = 0.16
Recall = 0.40
F-Measure = 0.23
Accuracy = 0.74

Another very interesting point on this chart is the spike at the end, on the 28th January 2012. On this day, news reports are focused on the match between Manchester United and Liverpool, the first between the two teams since the players from either side were involved in the racism-related event. There is a large spike in 'tension' and 'some tension' on this day and it is precisely this kind of tension indicator that first motivated this work — detecting spikes in tension that could lead to socially significant problems (i.e. at a football match).

Image 2 illustrates the tension over time as detected/classified by the COSMOS tension engine. Ideally, this should match Image 1 as closely as possible to demonstrate the engine matching the manually coded tension levels. It is most important that any spike in this graph matches the tension spike in Image 1 (20th Nov, 20th/21st Dec, and 27th Jan) and it is clear to see that these spikes do indeed exist in Image 2, demonstrating the engine detecting spikes that correlate with the human coders.

Finally, the experiment to plot the extreme differences in positive and negative sentiments on a daily basis produced Image 3. In Section 3.2.4 an assumption was made that a large daily differential between the mean maximum positive and negative sentiments would indicate a large difference in opinion, and therefore could indicate tension. That is, the bigger the difference between the two mean maximum scores (−5 to +5), the more tension exists, based on extreme opposite sentiments. For this assumption to be correct we would expect to see a large differential on the same days that tension spikes in Image 1 (i.e. when the manual coders identified raised levels of tension). Image 3 shows the maximum daily difference between positive and negative sentiments is 8 (3, −5–16th Nov/20th Dec), (4, −4–7th Dec/21st Dec/2nd Jan). Interestingly, three of these five dates follow the trend of spikes in tension as coded by the manual coders — once when Suarez was charged (16th Nov) and again when he was found guilty and banned (20/21st December).

On 7th December Suarez was charged with improper conduct by the Football Association due to a "one-fingered" salute to fans during a match. This has also brought a large differential in sentiment but as it is not related to the original racism event, the tweets were not manually coded as related to the event and therefore are not coded as containing tension. Nevertheless, tension clearly does exist on this day and looking back at Images 1 and 2, there is a clear spike in 'irrelevant' tweets. There does not appear to be any news story to support the 2nd Jan sentiment difference. Despite this anomaly, it does appear that while sentiment analysis did not perform as well as the tension engine at an individual tweet level, for a collection of tweets, a large differential in sentiment on a given day can indeed be used as a tension indicator.

## 5. Discussion & conclusion

The study has shown that human coders agree to an acceptable level that tension can be measurably observed in a corpus of tweets, and on different levels. A taxonomy emerged for the typology of a tweet where tweets can be passive-informative (simply passing on information), opinionated (often containing some level of tension), and aggressive (containing 'high' levels of tension).

While the results of recent academic studies provide some interesting insights into the "online" society, it must be recognized that the online society is a subset of wider society. It has been shown that internet access (in the UK) is lower than the national average amongst socially disadvantaged groups [35] and because of this, research data from the online community is most likely to be published by socially advantaged groups. However, more and more users are gaining access to the internet via mobile platforms and smartphone 'apps', which are broadening the demographic of online society and supporting a minute-by-minute commentary on daily life. We suggest that those responsible for ensuring societal cohesion and order must have the appropriate digital analysis methods available for them to capture, analyze and visualize the vast amount of data being published by online society.

Police recommendations following recent riots and protests have requested the ability to gather intelligence and analyze it

**Table 7**
Tension engine performance against human coded "tension" indicators.

| "Some Tension" | | Human Coders | |
|---|---|---|---|
| Machine | | Yes | No |
| | Yes | 163 | 40 |
| | No | 143 | 140 |

Precision = 0.80
Recall = 0.53
F-Measure = 0.64
Accuracy = 0.62

| "Tension" | | Human Coders | |
|---|---|---|---|
| Machine | | Yes | No |
| | Yes | 65 | 54 |
| | No | 67 | 300 |

Precision = 0.55
Recall = 0.49
F-Measure = 0.52
Accuracy = 0.75

| "High Tension" | | Human Coders | |
|---|---|---|---|
| Machine | | Yes | No |
| | Yes | 26 | 9 |
| | No | 22 | 429 |

Precision = 0.74
Recall = 0.54
F-Measure = 0.63
Accuracy = 0.94

in real-time in order that impending social unrest can be managed [27]. This paper focuses on refining the classification performance of the tension engine, however, with the constant stream of data being produced by the general public in the Twitter-sphere, and more people joining social networks on a daily basis, the COSMOS platform provides a tool for researchers observing events to conduct event-specific experiments where tension can be observed and visualized over time (and in real time), providing the potential to forecast social disruption via deviations from "normal" levels that are established over time (be that minutes, hours or days).

However, we must accept certain limitations in the application of COSMOS. The Twitter streaming API only supplies 1% of the total posts to Twitter for free. This means that we are missing 99% of the online conversations. However, 1% actually constitutes 3 million messages/tweets per day so the corpus is still significantly large, and we have shown that offline events are reflected in conversational fluctuations in the 1% sample. Indeed, if researchers were allowed access to 100% of the tweets posted to Twitter it would be necessary to conduct a scalability analysis to determine if handling data of this size would be possible. Based on the 3 million per day metric, researchers would need to process around 300 million tweets per day. If they wanted to store the data for later use the sheer amount of storage this data would need is far beyond the facilities of most academic researchers, and probably most public and private sector organizations.

Users of Twitter also morph the English language to fit their message into the message size limit of 140 characters, which makes it difficult to detect words and terms from specialist lexicons. For instance, while slangs such as "WTF", "LOL" and "LMAO", and abbreviations (e.g. 4 (for), u (you), and y (why)) may have informed classification tasks as unigram features, we have not yet included them in the lexicons of expletives for the tension classification engine. Furthermore, some tweets are posted in different languages other than English so we are unable to classify them. These factors also contribute to the variable quality and accuracy of Twitter data as a research tool.

Nonetheless, the results of the experiments conducted in this study show that a framework focused around lexicons of topic-specific actors, actions, accusations, and abusive and expletive terms can identify high level tension and distinguish those from lower levels, such that spikes in tension can be visualized over time. Sentiment analysis is also a useful tool to apply to a collection of tweets where the difference between the mean maximum positive and negative sentiments on a particular day can indicate tension through extreme differences in opinion and polarized views. As annotated dataset sizes improve, it may be interesting to test the performance of a combination of the three approaches with different weightings for each method on the basis of the best performing method at each point on the tension scale.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.techfore.2013.04.013.

## Acknowledgments

This work is based on the project Digital Social Research Tools, Tension Indicators and Safer Communities: A Demonstration of the Cardiff Online Social Media ObServatory (COSMOS) which was funded by the Economic and Social Research Council under the Digital Social Research Demonstrator Programme (Grant Reference: ES/J009903/1) and COSMOS: Supporting Empirical Social Scientific Research With a Virtual Research Environment which was funded by the Joint Information Systems Committee (JISC) under the Digital Infrastructure Research Programme, research tool strand.

## References

[1] Dyfed Powys Police, Community Tension Force Policy Document, 2008.
[2] J. Bollen, B. Gon√ßalves, G. Ruan, H. Mao, Happiness is assortative in online social networks, Artif. Life 17 (2011) 237–251.
[3] M. Thelwall, K. Buckley, G. Paltogou, Sentiment in Twitter events, J. Am. Soc. Inf. Sci. Technol. 62 (2011) 406–418.
[4] A. Tumasjan, T. Sprenger, P. Sandner, I. Welpe, Predicting elections with Twitter: what 140 characters reveal about political sentiment, International AAAI Conference on Weblogs and Social Media, Washington, D.C., 2010.
[5] J. Bollen, A. Pepe, H. Mao, Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena, Fifth International AAAI Conference on Weblogs and Social Media (ICWSM), Barcelona, 2009.
[6] B. Liu, Sentiment Analysis and Opinion Mining, Morgan & Claypool, 2012.
[7] P. Burnap, O. Rana, N. Avis, Making sense of self reported socially significant data using computational methods, Int. J. Soc. Res. Methodol. 16 (2013).
[8] R. Procter, A. Voss, F. Vis, How riot rumours spread on Twitter, Reading the Riots, Guardian.co.uk, 2011.
[9] A. Choudhary, W. Hendrix, K. Lee, D. Palsetia, W. Liao, Social media evolution of the Egyptian revolution, Commun. ACM 55 (2012) 74–80.
[10] G. Lotan, E. Graeff, M. Ananny, D. Gaffney, I. Pearce, D. Boyd, The revolutions were tweeted: information flows during the 2011 Tunisian and Egyptian revolutions, Int. J. Commun. 5 (2011) 1375–1405.
[11] D. Rathi, L. Given, Research 2.0: framework for qualitative and quantitative research in Web 2.0 environments, 43rd International Conference on System Sciences, 2010.
[12] P. Anderson, What is Web 2.0? Ideas, Technologies and Implications for Education, 2007.
[13] M. Thelwall, Social networks, gender and friending: An analysis of MySpace member profiles, J. Am. Soc. Inf. Sci. Technol. 59 (2008) 1321–1330.
[14] M. Thelwall, D. Wilkinson, S. Uppal, Data mining emotion in social network communication: gender differences in MySpace, J. Am. Soc. Inf. Sci. Technol. 61 (2010) 190–199.
[15] C. Fonio, F. Giglietto, R. Pruno, L. Rossi, S. Pedriolo, Eyes on you: analyzing user generated content for social science, Presented at Towards a Social Science of Web 2.0, 2007, (York, UK).
[16] J. Ahktar, S. Soria, Sentiment Analysis: Facebook Status Messages, Stanford University, 2009.
[17] S. Asur, B.A. Huberman, Predicting the Future With Social Media, , 2010.
[18] A. Pak, P. Paroubek, Twitter as a corpus for sentiment analysis and opinion mining, Seventh Conference on International Language Resources and Evaluation, 2010.
[19] A. Go, L. Huang, R. Bhayani, Twitter sentiment analysis, Final Projects from CS224N for Spring 2008/2009, Stanford Natural Language Processing Group, 2009.
[20] J. Read, Sing Emoticons to Reduce Dependency in Machine Learning Techniques for Sentiment Classification, ACL Student Research Workshop, Association for Computational Linguistics, Ann Arbor, Michigan, 2005.
[21] M. Thelwall, K. Buckley, G. Paltogou, D. Cai, A. Kappas, Sentiment strength detection in short informal text, J. Am. Soc. Inf. Sci. Technol. 61 (2010) 25442558.
[22] C. Yang, K. Hsin-Yih Lin, H. Chen, Emotion classification using web blog corpora, IEEE/WIC/ACM International Conference on Web Intelligence, IEEE Computer Society, Washington, DC, USA, 2007.
[23] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up? Sentiment classification using machine learning techniques, Empirical Methods in Natural Language Processing, 2002.
[24] K. Dave, S. Lawrence, D. Pennock, Mining the peanut gallery: opinion extraction and semantic classification of product reviews, 12thlinternational conference on World Wide Web, ACM, New York, NY, USA, 2003.
[25] L. Barbosa, J. Feng, Robust sentiment detection on Twitter from biased and noisy data, 23rd International Conference on Computational Linguistics, Association for Computational Linguistics, 2010.
[26] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede, Lexicon-based methods for sentiment analysis, Comput. Linguist. 37 (2011) 267–307.
[27] HMIC, Policing Public Order, An Overview and Review of Progress Against the Recommendations of Adapting to Protest and Nurturing the British Model of Policing, 2011.

[28] A. Bruns, Y. Liang, Tools and methods for capturing Twitter data during natural disasters, First Monday (2012).
[29] K. Krippendorff, Computing Krippendorff's Alpha-Reliability, 2005.
[30] H. Sacks, Lectures on Conversation, Vol I and II, Basil Blackwell, Oxford, 1992.
[31] W. Housley, The moral discrepancy device and 'fudging the issue' in a political radio news interview, Sociology 34 (2002).
[32] W. Housley, R. Fitzgerald, The reconsidered model of membership categorisation analysis, Qual. Res. 2 (2002).
[33] W. Housley, R. Fitzgerald, Membership categorization, culture and norms-in-action, Discourse Soc. 20 (2009).
[34] J. Brooks, A. Rawls, Steps toward a socio-technical categorization scheme for communication and information standards, iConference, 2012, pp. 407–414.
[35] S. Coleman, E. Normann, New Media and Social Inclusion, Hansard Society, 2000.
[36] A. Edwards, W. Housley, L. Sloan, M.L. Williams, M. Williams, Digital Social Research and the Sociological Imagination: Surrogacy, Augmentation and Re-orientation, International Journal of Social Research Methodology 16 (3) (2013).
[37] M.L. Williams, A. Edwards, W. Housley, P. Burnap, O. Rana, N. Avis, J. Morgan, L. Sloan, Policing cyber-neighbourhoods: Tension monitoring and social media networks, Policing and Society (2013).

**Pete Burnap** is a lecturer at Cardiff School of Computer Science & Informatics. He is an inter-disciplinary computer scientist currently working on funded research projects with social science, engineering and architecture. His expertise covers distributed computing, information and network security, probabilistic risk modeling and data analysis/machine learning. He is the PI on the COSMOS JISC project focusing on computational social informatics using text mining and network analysis techniques to address social science research questions. He is a Co-I on the WEFO funded SOLCER project investigating probabilistic models for low carbon energy systems' risk, efficiency and effectiveness.

**Omer Rana** is a professor of Performance Engineering at Cardiff School of Computer Science & Informatics and formerly deputy director of the Welsh eScience Centre. His expertise covers high performance distributed computing, multi-agent systems and data mining/analysis. He is a Co-I on the ESRC COSMOS and JISC projects focusing on social media analysis and mining. He has published over 200 papers in international journals and refereed conferences/workshops. He holds a PhD in Neural Computing and Parallel Architectures from Imperial College, London. He was involved in a theme on "Distributed, Dynamic, Data Intensive Applications" at the National eScience Institute (co-funded by the National Science Foundation in the US), focusing on "big-data" in computational science.

**Matthew Williams** is a senior lecturer in Criminology at the School of Social Sciences, Cardiff University, UK. He has published widely in the fields of cybercrime, digital social research methodology and diversity and criminal justice. Currently, he is the principal investigator on the ESRC project 'Digital Social Research Tools, Tension Indicators and Safer Communities: A demonstration of the Cardiff Online Social Media ObServatory (COSMOS)' REF:ES/J009903/1 (2011–2013) and the 'All Wales Hate Crime Project', funded by the Big Lottery Fund, UK (2010–2013). He is co-investigator on the JISC project 'Supporting Empirical Social Science Research with a Virtual Research Environment' (2012–2013) and an ESRC DSR Community Fund project entitled 'Requirements Analysis for Social Media Analysis Research Tools' (2012–2013).

**William Housley** is a reader in Sociology at the School of Social Sciences, Cardiff University, Wales, UK. He has published widely in the fields of interactionism, ethnomethodology, membership categorization analysis, social research methods, ethnography and computational sociology. His substantive research areas include team communication and decision making in organizations, the sociology of making, computational social science and mediated political and policy communications.

**Adam Edwards** is a senior lecturer in criminology in the School of Social Sciences at Cardiff University, UK. He is also director of the Cardiff research Centre for Crime, Law and Justice. He has been a member of the European Society of Criminology since its inception in 2000 and has directed two of its working groups, the 'European Governance of Public Safety Research Network' (2003–2009) and the 'Crime, Science and Politics' group (2009–2011, see: http://www.esc-eurocrim.org/workgroups.shtml#safety). His principal research interests are in the politics of public safety, the organization of serious crimes and the prospects for comparative European criminology. He is involved in research into the problems, responsibilities and expertise entailed in the management of urban security in Europe as part of 'project Urbis', which has been funded by the European Union's Leonardo Lifelong Learning Programme.

**Nick Avis** is a professor of Interactive Visualization and Virtual Environments at Cardiff School of Computer Science & Informatics. He is a Co-I for the ESRC COSMOS and JISC projects focusing on social media analysis and mining. His research interests include: interactive and real time visualization and virtual/augmented reality systems; computational steering; remote rendering; interactive grid middleware and multi-scale, multi-physics soft tissue representation and modelling.

**Luke Sloan** is a Lecturer in Quantitative Methods in the Cardiff University School of Social Sciences. His research interests include reconceptualising tradition notions of quantitative methodology for integration with social media and augmenting curated with naturally occurring data.

**Jeffrey Morgan** is a Research Associate in the Cardiff University School of Social Sciences. He is responsible for designing and implementing the COSMOS platform. His research interests include human-computer interaction and the design of highly-interactive information retrieval and visualisation systems.