Entity-based Opinion Mining from Text and Multimedia

Diana Maynard and Jonathon Hare

1 Introduction

Social web analysis is all about the users who are actively engaged and generate content. This content is dynamic, reflecting the societal and sentimental fluctuations of the authors as well as the ever-changing use of language. Social networks are pools of a wide range of articulation methods, from simple "Like" buttons to complete articles, their content representing the diversity of opinions of the public. User activities on social networking sites are often triggered by specific events and related entities (e.g. sports events, celebrations, crises, news articles) and topics (e.g. global warming, financial crisis, swine flu).

With the rapidly growing volume of resources on the Web, archiving this material becomes an important challenge. The notion of community memories extends traditional Web archives with related data from a variety of sources. In order to include this information, a semantically-aware and socially-driven preservation model is a natural way to go: the exploitation of Web 2.0 and the wisdom of crowds can make web archiving a more selective and meaning-based process. The analysis of social media can help archivists select material for inclusion, while social media mining can enrich archives, moving towards structured preservation around semantic categories. In this paper, we focus on the challenges in the development of opinion mining tools from both textual and multimedia content.

We focus on two very different domains: socially aware federated political archiving (realised by the national parliaments of Greece and Austria), and socially contextualized broadcaster web archiving (realised by two large multimedia broad-

Diana Maynard
Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello, Sheffield, S1 4DP, UK e-mail: diana@dcs.shef.ac.uk

Jonathon Hare
Electronics and Computer Science, University of Southampton, Southampton, Hampshire, SO17 1BJ, UK e-mail: jsh2@ecs.soton.ac.uk

casting organizations based in Germany: Sudwestrundfunk and Deutsche Welle). The aim is to help journalists and archivists answer questions such as what the opinions are on crucial social events, how they are distributed, how they have evolved, who the opinion leaders are, and what their impact and influence is.

Alongside natural language, a large number of the interactions which occur between social web participants include other media, in particular images. Determining whether a specific non-textual media item is performing as an opinion-forming device in some interaction becomes an important challenge, more so when the textual content of some interaction is small or has no strong sentiment. Attempting to determine a sentiment value for an image clearly presents great challenges, and this field of research is still in its infancy. We describe here some work we have been undertaking, firstly to attempt to provide a sentiment value from an image outside of any specific context, and secondly to utilise the multimodal nature of the social web to assist the sentiment analysis of either the multimedia or the text.

2 Related Work

While much work has recently focused on the analysis of social media in order to get a feel for what people think about current topics of interest, there are, however, still many challenges to be faced. State of the art opinion mining approaches that focus on product reviews and so on are not necessarily suitable for our task, partly because they typically operate within a single narrow domain, and partly because the target of the opinion is either known in advance or at least has a limited subset (e.g. film titles, product names, companies, political parties, etc.).

In general, sentiment detection techniques can be roughly divided into lexicon-based methods [1] and machine-learning methods, e.g. [2]. Lexicon-based methods rely on a sentiment lexicon, a collection of known and pre-compiled sentiment terms. Machine learning approaches make use of syntactic and/or linguistic features, and hybrid approaches are very common, with sentiment lexicons playing a key role in the majority of methods. For example, [3] establish the polarity of reviews by identifying the polarity of the adjectives that appear in them, with a reported accuracy of about 10% higher than pure machine learning techniques. However, such relatively successful techniques often fail when moved to new domains or text types, because they are inflexible regarding the ambiguity of sentiment terms. The context in which a term is used can change its meaning, particularly for adjectives in sentiment lexicons [4]. Several evaluations have shown the usefulness of contextual information [5], and have identified context words with a high impact on the polarity of ambiguous terms [6]. A further bottleneck is the time-consuming creation of these sentiment dictionaries, though solutions have been proposed in the form of crowdsourcing techniques[1].

[1] http://apps.facebook.com/sentiment-quiz

Almost all the work on opinion mining from Twitter has used machine learning techniques. [7] aimed to classify arbitrary tweets on the basis of positive, negative and neutral sentiment, constructing a simple binary classifier which used n-gram and POS features, and trained on instances which had been annotated according to the existence of positive and negative emoticons. Their approach has much in common with an earlier sentiment classifier constructed by [8], which also used unigrams, bigrams and POS tags, though the former demonstrated through analysis that the distribution of certain POS tags varies between positive and negative posts. One of the reasons for the relative paucity of linguistic techniques for opinion mining on social media is most likely due to the difficulties in using NLP on low quality text [9]; for example. the Stanford NER drops from 90.8% F1 to 45.88% when applied to a corpus of tweets [10].

There have been a number of recent works attempting to detect sarcasm in tweets and other user-generated content [11, 12, 13, 14], with accuracy typically around 70-80%. These mostly train over a set of tweets with the #sarcasm and/or #irony hashtags, but all simply try to classify whether a sentence or tweet is sarcastic or not (and occasionally, into a set of pre-defined sarcasm types). However, none of these approaches go beyond the initial classification step and thus cannot predict how the sarcasm will affect the sentiment expressed. This is one of the issues that we tackle in our work.

Extracting sentiment from images is still a research area that is in its infancy and not yet prolifically published. However, those published often use small datasets for their ground truth on which to build SVM classifiers. Evaluations show systems often respond only a little better than chance for trained emotions from general images [15]. The implication is that the feature selection for such classification is difficult. [16] used a set of colour features for classifying their small ground-truth dataset, also using SVMs, and publish an accuracy of around 87%. In our work, we expand this colour-based approach to use other features and also use the wisdom of the crowd for selecting a large ground-truth dataset.

Other papers have begun to hint at the multimodal nature of web-based image sentiment. Earlier work, such as [17], is concerned with similar multimodal image annotation, but not specifically for sentiment. They use latent semantic spaces for correlating image features and text in a single feature space. In this paper, we describe the work we have been undertaking in using text and images together to form sentiment for social media.

3 Opinion Mining from Text

3.1 Challenges

There are many challenges inherent in applying typical opinion mining and sentiment analysis techniques to social media. Microposts such as tweets are, in some

sense, the most challenging text type for text mining tools, and in particular for opinion mining, since the genre is noisy, documents have little context and assume much implicit knowledge, and utterances are often short. As such, conventional NLP tools typically do not perform well when faced with tweets [18], and their performance also negatively affects any following processing steps.

Ambiguity is a particular problem for tweets, since we cannot easily make use of coreference information: unlike in blog posts and comments, tweets do not typically follow a conversation thread, and appear much more in isolation from other tweets. They also exhibit much more language variation, and make frequent use of emoticons, abbreviations and hashtags, which can form an important part of the meaning. Typically, they also contain extensive use of irony and sarcasm, which are particularly difficult for a machine to detect. On the other hand, their terseness can also be beneficial in focusing the topics more explicitly: it is very rare for a single tweet to be related to more than one topic, which can thus aid disambiguation by emphasising situational relatedness.

In longer posts such as blogs, comments on news articles and so on, a further challenge is raised by the tracking of changing and conflicting interpretations in discussion threads. We investigate first steps towards a consistent model allowing for the pinpointing of opinion holders and targets within a thread (leveraging the information on relevant entities extracted).

We refer the reader to [18] for our work on twitter-specific IE, which we use as pre-processing for the opinion mining described below. It is not just tweets that are problematic, however; sarcasm and noisy language from other social media forms also have an impact. In the following section, we demonstrate some ways in which we deal with this.

3.2 Opinion Mining Application

Our approach is a rule-based one similar to that used by [1], focusing on building up a number of sub-components which all have an effect on the score and polarity of a sentiment. In contrast, however, our opinion mining component finds opinions relating to previously identified entities and events in the text. The core opinion mining component is described in [19], so we shall only give an overview here, and focus on some issues specific to social media which were not dealt with in that work, such as sarcasm detection and hashtag decomposition.

The detection of the actual opinion is performed via a number of different phases: detecting positive, negative and neutral words, identifying factual or opinionated versus questions or doubtful statements, identifying negatives, sarcasm and irony, analysing hashtags, and detecting extra-linguistic clues such as smileys. The application involves a set of grammars which create annotations on segments of text. The grammar rules use information from gazetteers combined with linguistic features (POS tags etc.) and contextual information to build up a set of annotations and features, which can be modified at any time by further rules. The set of gazetteer

lists contains useful clues and context words: for example, we have developed a gazetteer of affect/emotion words from WordNet [20]. The lists have been modified and extended manually to improve their quality.

Once sentiment words have been matched, we find a linguistic relation between these and an entity or event in the sentence or phrase. A Sentiment annotation is created for that entity or event, with features denoting the polarity (positive or negative) and the polarity score. Scores are based on the initial sentiment word score, and intensified or decreased by any modifiers such as swear words, adverbs, negation, sarcasm etc, as explained next.

Swear words are particularly prolific on Twitter, especially on topics such as popular culture, politics and religion, where people tend to have very strong views. To deal with these, we match against a gazetteer list of swear words and phrases, which was created manually from various lists found on the web and from manual inspection of the data, including some words acquired by collecting tweets with swear words as hashtags (which also often contain more swear words in the main text of the tweet).

Much useful sentiment information is contained within hashtags, but this is problematic to identify because hashtags typically contain multiple words within a single token, e.g. #notreally. If a hashtag is camelcased, we use the capitalisation information to create separate tokens. Second, if the hashtag is all lowercase or all uppercase, we try to form a token match against the Linux dictionary. Working from left to right, we look for the longest match against a known word, and then continue from the next offset. If a combination of matches can be found without a break, the individual components are converted to tokens. In our example, #notreally would be correctly identified as “not” + “really”. However, some hashtags are ambiguous: for example, ”#greatstart” gets split wrongly into the two tokens ”greats” + ”tart”. These problems are hard to deal with; in some cases, we could make use of contextual information to assist.

We conducted an experiment to measure the accuracy of hashtag decomposition, using a corpus of 1000 tweets randomly selected from the US elections crawl that we undertook in the project. 944 hashtags were detected in this corpus, of which 408 were identified as multiword hashtags (we included combinations of letters and numbers as multiword, but not abbreviations). 281 were camelcased and/or combinations of letters and nubers, 27 were foreign words, and the remaining 100 had no obvious token-distinguishing features. Evaluation on the hard-to-recognise cases (non-camel-cased multiword hashtags) produced scores of 86.91% Precision, 90% Recall, and an F-measure of 88.43%. Given that these hard-to-resolve combinations form roughly a quarter of the multiword hashtags in our corpus, and that we are entirely successful in decomposing the remaining hashtags, this means that the overall accuracy for hashtag decomposition is much higher.

In addition to using the sentiment information from these hashtags, we also collect new hashtags that typically indicate sarcasm, since often more than one sarcastic hashtag is used. For this, we used the GATE gazetteer list collector to collect pairs of hashtags where one was known to be sarcastic, and examined the second hashtag manually. From this we were able to identify a further set of sarcasm-indicating

hashtags, such as #thanksdude, #yay etc. Further investigation needs to be performed on these to check how frequently they actually indicate sarcasm when used on their own.

Finally, emoticons are processed like other sentiment-bearing words, according to another gazetteer list, if they occur in combination with an entity or event. For example, the tweet ”They all voted Tory :-(” would be annotated as negative with respect to the target ”Tory”. Otherwise, as for swear words, if a sentence contains a smiley but no other entity or event, the sentence gets annotated as sentiment-bearing, with the value of that of the smiley from the gazetteer list.

Once all the subcomponents have been run over the text, a final output is produced for each sentiment-bearing segment, with a polarity (positive or negative) and a score, based on combining the individual scores from the various components (for example, the negation component typically reverses the polarity, the adverbial component increases the strength of the sentiment, and so on. Aggregation of sentiment then takes place for all mentions of the same entity/event in a document, so that summaries can be created.

4 Mining images and their context

Images are often used to illustrate the opinions expressed by the text of a particular media item. By themselves, images also have the ability to convey and elicit opinions, emotions and sentiments. In order to investigate how images are used in the opinion formation process, we have been developing tools that allow in-depth analysis of specific elements within an image to be used to quantify elements of opinion and sentiment, and allow the reuse of images within an archive or corpus to be contextualised with respect to diverse time and opinion axes.

4.1 Challenges

The main challenge with annotating non-textual media is that the underlying tokens within it are considerably less explicit than in textual media. In images and video, these underlying tokens are groups of pixels (compared with groups of characters [words] in text). As well as having multiple dimensions, the tokens have considerably more variation when representing exactly the same concept, and so using dictionaries and other traditional text-based techniques often becomes impractical. State of the art computer vision and automated image understanding is still a relatively immature subject for most general applications. This “semantic gap” between what computer vision can achieve and the level of understanding required for tasks such as sentiment analysis is why extracting opinions from images is so difficult.

Even though computer vision is challenging, considerable advances have been made in recent years. This is in particular true for the detection and recognition of

certain types of objects or entities. In terms of the types of entities often recognised by Named Entity Recognition (NER) tasks in test documents, there are a number of relatively mature visual equivalents:

1. The detection of *Person* entities in images can be acheived in a fairly robust manner by detecting human faces, and face recognition technologies can help recognise and disambiguate the specific person. This is discussed in more detail in Section 4.2 for more information.
2. *Organisation* entities can be detected and recognised in images by looking for certain indicators, such as the logo of the organisation. Techniques that can robustly detect rigid patterns in images (such as logos) are common-place in modern computer vision (e.g. [21, 22]).
3. The recognition of *Place* entities is currently a hot topic in the multimedia analysis community, and a number of techniques for determining the where an image was taken have been proposed. From a purely visual analysis point of view, these techniques tend to either work by direcly matching the image against large datasets of images with known locations (which tends to only work well for well-known places), or by estimating visual attributes that can help infer location (for example, that a photo depicts a beach scene, thus limiting the possible locations to coast lines). The former techniques tend to have very high precision, but low recall, whereas the latter techniques have much less precision (but higher recall).

One big challenge with all these approaches to extracting entities from images is dealing with the shear amount of data required. For all the techniques, large amounts of image data is required to learn the visual representations. In some cases, this makes the problem intractable without additional constraints. For example, in the case of face recognition, or even logo recognition, it is not possible to have multiple images of all the people (or logos) in the world from which to train discriminative classifiers. Typically, these problems are constrained by deciding apriori specifically who (or what) needs to be detected in the images being analysed. Another way of constraining the analysis is to make use of the any available information from the context of the image in question (for example analysis of surrounding text, titles, tags, etc.), and use this to guide the visual analysis.

4.2 Exploring human faces

Human faces are an obvious starting point for image analysis as they can potentially tell us who is in the image, as well as allowing us to make inferences to that persons emotional state. Before any higher-level analysis can occir, faces must first be detected in an image. The problem of face detection has been studied for a very long time in the computer vision field, and whilst not solved completetly has a number of acceptable solutions (under certain constraints, such as requiring the face be "frontal" or approximately facing the camera).

Whilst by no means the only (or best) approach, the algorithm for face detection developed by Viola and Jones [23] is probably the most widespread computer-vision technique of all time. Viola and Jones's technique works by training cascades of simple classifiers based on certain small patterns of light and dark pixels (these patterns are often referred to as Haar-like features, as they approximate the Haar wavelet function). When trained on large sets of human face images, the resultant classifier cascade can detect faces in images robustly and efficiently. In the case of human faces, the trained classifiers recognise patterns common to all faces, such as the areas directly above an below the eye generally having lighter intensity than the eye itself.

4.2.1 Analysing facial expression

Once a face has been localised it is possible to make measurable estimates of that individuals' facial expression in the image [24, 25, 26], as well as other attributes such as gender. Facial expressions are of particular interest because psychological studies have shown facial expressions can be used to infer the emotional state of the individual [27], and thus be used to infer sentiment. The Facial Action Coding System [28] (FACS) is a tool developed by psychologists to provide a standardised way of describing the expressions of faces. Codes represent muscular actions in the face (such as "inner eyebrow raising", or "lip corner puller"). Further coding systems such as EMFACS [29] and FACSAID [30] provide combinations of FACS codes that represent emotions (for example, activation of the lip corner puller AU6 and the cheek raiser AU12 actions imply happiness).

Given a detection of a human face in an image, it is possible to fit a flexible *shape* model that describes the overall intrinsic characteristics of the depicted individuals' face and their expression, as well as extrinsic characteristics such as the pose of the person relative the camera. Active Shape Models [31] (ASMs), Active Appearance Models [32] (AAMs) and Constrained Local Models [33] (CLMs) are well-studied algorithms for fitting a flexible shape to an image using the image's content to choose the best position for the vertices of the shape whilst constraining the shape to be plausible (based on a set of training examples that define the extents of the shape). As these models are both parametric (the shape is controlled by a small number of parameters) and generative (they allow a face to be reconstructed using these parameters), a large range of poses, expressions and appearances (skin textures) can be generated. Fitting a model to an image is a constrained optimisation problem in which the parameters of the model are iteratively updated in order to minimise the difference between the generated model and the image. Once a model is fitted to an image, the parameters can then be used as input to an expression classifier that can determine an expression label for the face. More specifically, the muscular movements encoded by FACS map to combinations of parameters in the face model, so a classifier can be potentially trained to recognise these actions [34, 35, 36]. Figure 1 shows a screenshot of our experimental CLM-based expression recognition system which has been trained to recognise FACS AUs in a

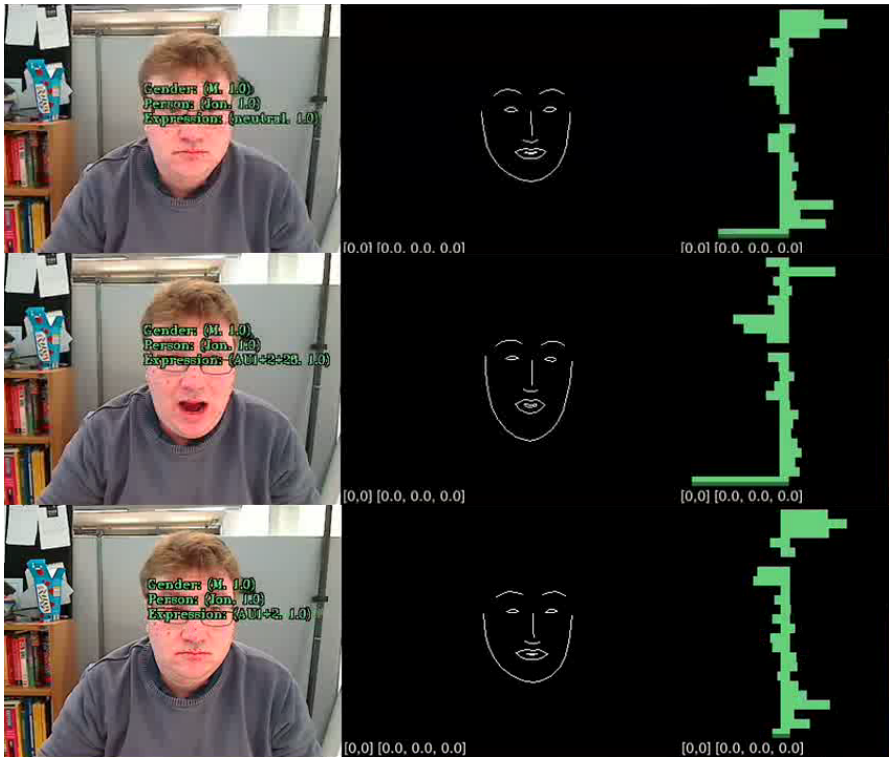

Fig. 1 Recognition of expressions in a laboratory setting using a CLM. The bars on the right illustrate the values of the parameter vector which define the shape of the model shown in the centre. Automated fitting techniques are used to adjust the values in the parameter vector so that the generated shape optimally matches the face in the image on the left.

laboratory setting, with highly constrained imaging conditions (i.e. restricted pose, uncluttered background, etc).

Unfortunately, training a system to detect the full set of action units required for the different emotional states is difficult due to the lack of publicly available data. A second problem directly relates to the facial models themselves, in that it is quite difficult to build a shape model (ASM, AAM or CLM) that will accurately fit all faces, which is essential for the accurate measurement of the shape parameters needed for expression classification. A third and final problem is that accurate detection of a face is required to initialise the fitting of the model; whilst face detection techniques are quite mature, they can still have major problems working in real-world images where the faces are not exactly frontal to the camera, or there are shadows or contrast issues. Using real-world images collected from the web and social web, we found that inaccuracies in the face model alignment would regularly cause misclassification of the action units, and therefore the expressions. Figure 2 shows some examples of the trained CLM model illustrated in Figure 1 applied to example

images collected from social media that are related to the US Elections. Notice in particular how poorly the model fits to Michelle Obama's face (and causes the misclassification of gender as a side effect). As this is a rapidly moving area of research, it will be interesting to see how expression modelling techniques develop over the coming years, especially in the presence of benchmarks such as Facial Expressions in the Wild[2] [37].

4.2.2 Recognising People

Once faces have been detected, recent advancements in face recognition means that people can be recognised with relatively high accuracy from within a small search space (i.e., relatively a small set of people to choose from). The problem with a general media analysis scenario is that the search space is effectively infinite, and current face recognition algorithms tend to deteriorate rapidly as the search space gets larger. One option that we have started to explore in our recent work is to apply entity recognition to any available contextual text to extract mentions of people, which we then use constrain the face recogniser's search space to a small subset of person entities. For well-known personalities and people whose photos can be found on the Internet, a web-based image search can be used to automatically retrieve example images of those people from which a face recognition algorithm can be trained [38]. An illustration of the overall process used in our recent experiments is shown in Figure 3. This overall process of using the contextual information to guide what to look for in the image is equally applicable to other types of entity, such as organisations with their corporate logos.

4.3 Contextualising image reuse

One way of gathering interesting insights into the social web is to look at how media spreads. In particular, we an measure how it is reused and talked about over time, and whether the aspects of the context, such as sentiment, change. One very powerful affordance gained from using near-duplicate images in this way is it that the analysis is agnostic of the context, and in particular can be used to link together very different contexts which share the same image. From a practical point of view, duplicate images can be used to infer links between social media documents with text in the variety of human languages without the need to explicitly understand those languages.

Detecting duplicate images is not just a matter of looking at the url from which the media is hosted because the same image is often hosted in many different locations, often with subtle (or not so subtle) changes from compression, cropping, rotation, etc. Using recent computer vision techniques, *near-duplicate* images can

[2] http://cs.anu.edu.au/few

Fig. 2 Examples illustrating a CLM-based shape model with associated attribute classifiers applied to real images from the social web.

be detected efficiently across very large static datasets [39, 40] and streams of (social) media [41]. The technology behind these systems varies, but typically relies on some form of robust image feature extraction followed by an indexing step to enable images to be efficiently compared. The SIFT local feature [42] is a popular choice to describe the image's content as it is highly robust to the typical transformations that make images near-duplicates rather than exact duplicates. For the indexing step,

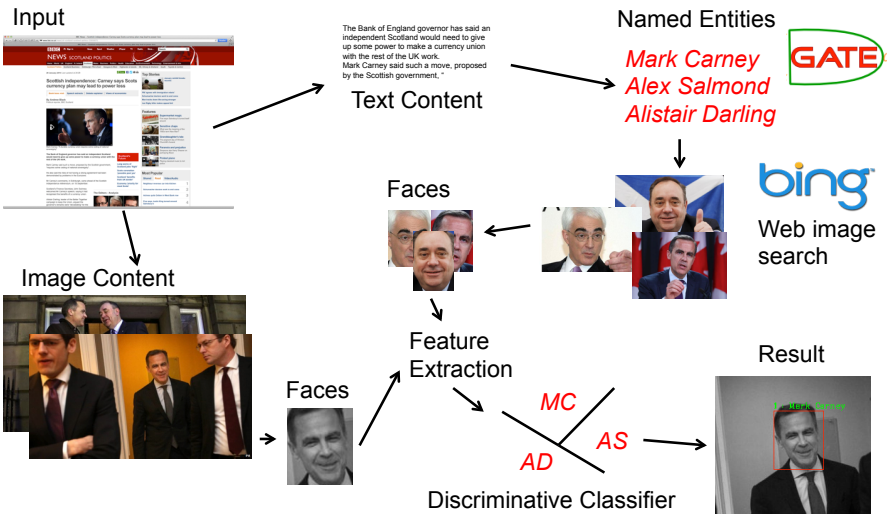

Fig. 3 Automated face verification using the names of people detected from the contextual text.

vector-quantisation followed by storage in an inverted index [40, 43], and locality sensitive hashing [44, 39, 41] are popular approaches.

4.3.1 Mining temporal reuse

Given a corpus of documents containing images in which we know the time that the document was created or posted, we can start to explore how a given image is reused over time. Figure 4 shows a screenshot of an experimental visualisation that displays duplicate images on a timelines, based on the date of the document that contained the image. From this visualisation it is possible to see how the incidence of the image varies over time as well as identifying clusters which may signify important time periods within the narrative of the image. In particular, in the specific case of the data used for the visualisation in Figure 4 (which in this case was created from a web crawl) it is possible to see hidden patterns of reuse being exposed. The topmost band shows images of a painting called "The Scream" by Edvard Munch. In 2004 this painting was stolen from a museum in Norway and it is here where the image is first used. During the following 3 years, the story about the stolen painting appeared in news articles as the thieves were arrested and charged, and the painting then recovered; three separate events in the narrative of this story which are elucidated by the visualisation.

Interestingly, the example shown in Figure 4 also displays a time correlation between the picture of The Scream and the picture two lines below. This second picture is a photograph of the thieves making off with the painting itself. This correlation can be investigated by looking at the contextual information from the document in

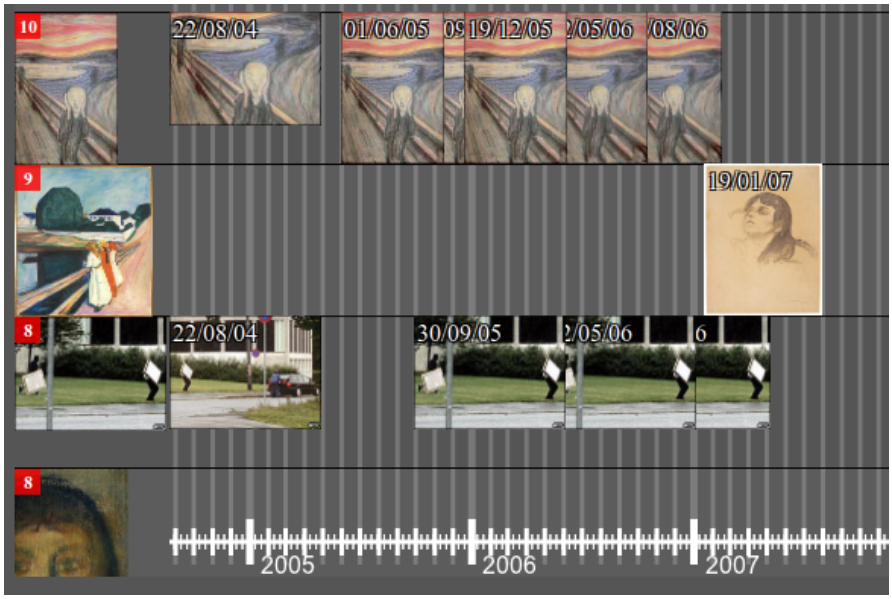

Fig. 4 Visualising how images are reused over time.

which the image was embedded; in this case, the correlation is, perhaps, expected as the photograph is related to the story of the stolen painting. However, the stories to which that photograph is related are very different to those to which the picture of the painting are related, despite the correlation. Indeed, examining the narrative thread exposed by the visualisation makes it clear that the picture of the painting is associated with the narrative of the painting being stolen, whereas the photograph of the thieves is associated with complementary articles about protecting museum artefacts.

4.3.2 Mining sentiment and opinion polarities from reused images

Recently, we developed a system called Twitter's Visual Pulse [41] which finds near-duplicate images over fixed time periods in a live Twitter stream. By extracting the sentiment from the tweets associated with these duplicate images (using the techniques in Section 3), we can find out how the image is used in different contexts. In many cases, the image may be reused in contexts which are, overall, sentimentally ambivalent; however, there may be cases where an image is used in a consistent way - for example, a particular image may be used in consistently positive tweets. We form a discrete probability distribution for images falling in specific sentiment categories, which we can use to assign sentiment probabilities to the image when it is further reused, particularly in cases where the textual sentiment analysis is inconclusive. When a context has conflicting opinions, or an opinion is not evident, then the

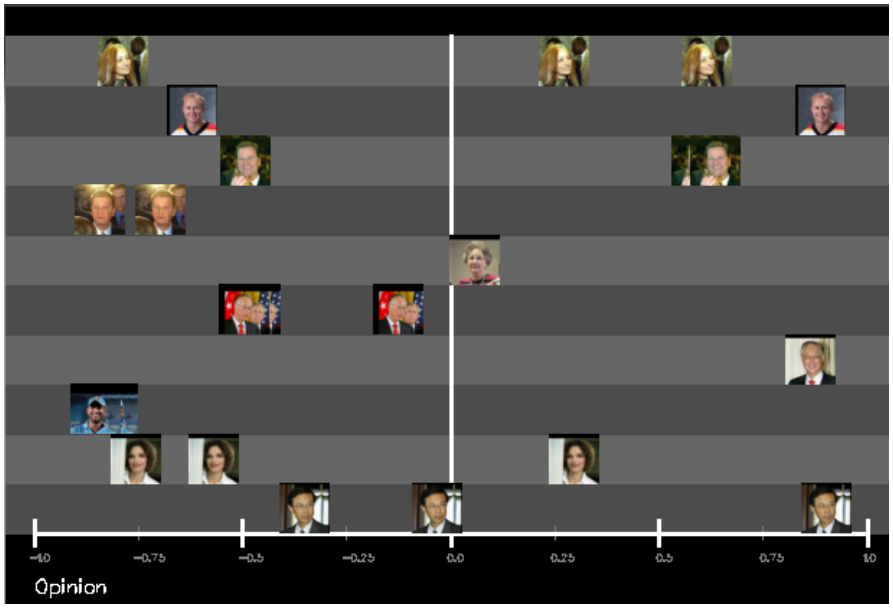

Fig. 5 Visualising how reused images vary with respect to the opinion polarity of their context.

image may be able to provide clues as to the article's sentiment: if it contains an image which has been reused many times in articles that have particular opinions, the ambiguous article can be associated with that opinion through the association with the image. Because the image matching is purely visual, this technique will work across language barriers, such that articles in a language that cannot be analysed could still have sentiment scores associated with them.

It can also be instructive to visualise the sentiments or opinion polarities associated with the contexts of particular images as they are reused. Figure 5 shows an example of this.

4.4 Exploring multimodal sentiment, privacy and attractiveness in social images

Opinion and sentiment are rather complex notions that can be very difficult to predict purely from visual data alone. A more fruitful approach is to consider the image (or other media modality) in the context in which it appears, whether that be an image on Flickr or video on YouTube surrounded by tags and comments provided by humans; or an image in a news item surrounded by the text of the article to which it relates. State-of-the-art research on the sentiment analysis of images (see e.g. [15, 45, 46, 16, 47]) has already begun to explore how the analysis of textual content and the analysis of visual content can complement each other. Recently, we

have been exploring how visual content and contextual information can be leveraged to train machines to predict facets related to opinion formation.

4.4.1 Image sentiment

In less constrained multimedia, we cannot rely on there being faces in the images, and sentiment may be carried by other visual traits. Indeed, images may intrinsically have sentiment associated with them through design (say a poster for a horror film) or through association with a specific subject matter which may be context sensitive (say a photo of wind generators in the context of climate change). For these situations there are no specific algorithms we can use for extracting the sentiment. However, we can look for correlations between visual features and textual labels using classifiers and regressors trained over ground-truth datasets. Unfortunately, large, well labelled datasets for image sentiment are hard to come by. For that reason, we turned to user-provided image annotations to generate a large dataset to use for classification. Using SentiWordNet [48], we queried Flickr for the words that had the strongest positive and negative sentiments, and retrieved sets of images for each of them. Combined, these formed a ground-truth for positive and negative sentiment images. Full details of the dataset and the trained classifiers are described in [46], but we will summarise the conclusions here.

We gathered images for the 1000 strongest sentiment words from SentiWordNet. This resulted in 586,000 images, most of which had a resolution of more than 1 megapixel. We extracted global and local colour features (these describe the colour distribution in the image, and in the case of the local variant, a coarse spatial layout of the colour distribution) and SIFT local features [42] (which describe small patches of texture/pattern in the image) from the images. Using these features a linear SVM classifier was trained to recognise positive/negative sentiment. We observed that for small recall values, precision values of up to 70% can be reached. Due to the challenging character of this task, for high recall values, the precision degrades down to the random baseline. Interestingly, using mutual information, we were able to reverse engineer the correlations in the classifier to determine which features were correlated to which labels. We found that positive images had overall warm colours (reds, oranges, yellows, skin tones) and negative images had colder colours (blues, dark greens). The location of the colour had no real significance. The negative SIFT features seem dominated by a very light central blob surrounded by a much darker background, while the positive SIFT features are dominated by a dark blob on the side of the patch.

4.4.2 Image privacy

In terms of privacy classification, we have been able to construct classifiers using textual tags and visual features, both combined and separately, in order to predict whether an image is potentially of a private nature. This is directly related to opin-

ion formation, because it can potentially be used to identify images such as paparazzi shots and leaked private images which have been published or posted in public places.

For our privacy classification experiments [49, 50], we created a dataset of 90,000 "recently uploaded" images from Flickr with a minimum of 5 English tags. In order to create ground-truth, we created a social annotation game and used crowdsourcing to get the opinions of multiple individuals. In the game, users were able to select three different options for each image they were presented with: private, undecidable or public. Users were given the following advice before commencing the game: *Private photos are photos which have to do with the private sphere (like self portraits, family, friends, your home) or contain objects that you would not share with the entire world (like a private email). The rest is public. In case no decision can be made, the picture should be marked as undecidable.*

Altogether the participants annotated 83,820 images. Analysis showed that around 78% of photos were labeled as public or undecidable by all of the participating judges. This is to be expected due to the nature of images on Flickr, which are on the whole posted to be shared with the public at large. From the remaining 22% of photos, 12% were labeled as "private" by all the judges, and 10% received "private" votes from at least one of the judges. A subsample of the data with the highest annotator agreement was selected for performing classification experiments.

A selection of different visual features were extracted from the images for training input to linear SVM classifiers. The textual feature was a simple word-occurrence histogram, with stemming applied to the tags to reduce variability and group similar tags. The classifiers were created and evaluated for each individual feature, all visual features combined, and text and visual features combined. Combined features worked better than individual features; evaluation using precision-recall metrics showed a break-even point of 0.74 for visual features, 0.78 for textual features and 0.80 for combined text and visual features.

4.4.3 Image attractiveness

When considered within the context of the article or post in which it appears, we hypothesise that the attractiveness of a photograph can be a strong indicator of the opinion and sentiment expressed by the article. Currently, we are only beginning to scratch the surface of this area, but we have been investigating building computational models of attractiveness that take into account both visual features as well as surrounding contextual tags [51].

On the assumption that on Flickr, more attractive or aesthetically pleasing photographs have higher numbers of favourite assignments, we built a dataset of 400,000 images as follows: We randomly selected time periods of 20 minutes from a time span of 5 years 2005-2010. From each of the periods we selected at most 5 pictures from Flickr with the highest number of favourite assignments as positive examples, as well as the same number of photos without favourite assignments as

negative examples. We stopped after obtaining a set of 200,000 photos from each class.

Even though aesthetic and artistic quality cannot be quantitatively computed, it has been shown that certain visual features of images have significant correlation with them. For instance, appealing images tend to have higher colourfulness, increased contrast and sharpness [52]; we apply image analysis to extract these features. Bag-of-words textual features extracted from the title and tags can also provide information about the image quality and aesthetics. By training linear SVM classifiers, we are able to generate predictive models of image attractiveness using these features. Experiments (see [51] for full details) have shown that our visual features can provide reasonable performance (break-even-point of 0.67 with respect to the precision-recall curve), whilst combinations of the textual and visual features perform better than either the textual or visual features alone (combined feature break-even-point of 0.84).

5 Conclusions

In this paper, we have described the general approach we undertake to the analysis of social media, using a combination of textual and multimedia opinion mining tools. It is clear that both opinion mining in general, and the wider analysis of social media, are difficult tasks from both perspectives, and there are many unresolved issues. The modular nature of our approach also lends itself to new advances in a range of subtasks: from the difficulties of analysing the noisy forms of language inherent in tweets, to the problems of dealing with sarcasm in social media, to the ambiguities inherent in such forms of web content that inhibit both textual and multimedia analysis tools. Furthermore, to our knowledge this is the first system that attempts to combine such kinds of textual and multimedia analysis tools in an integrated system, and preliminary results are very promising, though this is nevertheless very much ongoing research. Future work includes further development of the opinion mining tools: we have already begun investigations into issues such as sarcasm detection, more intricate use of discourse analysis and so on.

Acknowledgements This work was supported by the European Union under grant agreements No. 270239 Arcomem[3] and No. 610829 DecarboNet[4].

References

1. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Computational Linguistics **1** (2011) 1–41

[3] http://www.arcomem.eu

[4] http://www.decarbonet.eu

2. Boiy, E., Moens, M.F.: A machine learning approach to sentiment analysis in multilingual web texts. Information Retrieval **12** (2009) 526–558
3. Moghaddam, S., Popowich, F.: Opinion polarity identification through adjectives. CoRR **abs/1011.4623** (2010)
4. Mullaly, A., Gagné, C., Spalding, T., Marchak, K.: Examining ambiguous adjectives in adjective-noun phrases: Evidence for representation as a shared core-meaning. The Mental Lexicon **5** (2010) 87–114
5. Weichselbraun, A., Gindl, S., Scharl, A.: A context-dependent supervised learning approach to sentiment detection in large textual databases. Journal of Information and Data Management **1** (2010) 329–342
6. Gindl, S., Weichselbraun, A., Scharl, A.: Cross-domain contextualisation of sentiment lexicons. In: Proceedings of 19th European Conference on Artificial Intelligence (ECAI-2010). (2010) 771–776
7. Pak, A., Paroubek, P.: Twitter Based System: Using Twitter for Disambiguating Sentiment Ambiguous Adjectives. In: Proceedings of the 5th International Workshop on Semantic Evaluation. (2010) 436–439
8. Go, A., Bhayani, R., , Huang, L.: Twitter sentiment classification using distant supervision. Technical Report CS224N Project Report, Stanford University (2009)
9. Derczynski, L., Maynard, D., Aswani, N., Bontcheva, K.: Microblog-Genre Noise and Impact on Semantic Annotation Accuracy. In: Proceedings of the 24th ACM Conference on Hypertext and Social Media, ACM (2013)
10. Liu, X., Zhang, S., Wei, F., Zhou, M.: Recognizing named entities in tweets. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. (2011) 359–367
11. Tsur, O., Davidov, D., Rappoport, A.: Icwsm–a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. (2010) 162–169
12. Liebrecht, C., Kunneman, F., van den Bosch, A.: The perfect solution for detecting sarcasm in tweets# not. WASSA 2013 (2013) 29
13. Reyes, A., Rosso, P., Veale, T.: A multidimensional approach for detecting irony in Twitter. Language Resources and Evaluation (2013) 1–30
14. Davidov, D., Tsur, O., Rappoport, A.: Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. In: Proceedings of the Fourteenth Conference on Computational Natural Language Learning, Association for Computational Linguistics (2010) 107–116
15. Yanulevskaya, V., Van Gemert, J., Roth, K., Herbold, A.K., Sebe, N., Geusebroek, J.M.: Emotional valence categorization using holistic image features. In: Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on. (2008) 101–104
16. Wei-ning, W., Ying-lin, Y., Sheng-ming, J.: Image retrieval by emotional semantics: A study of emotional space and feature extraction. In: Systems, Man and Cybernetics, 2006. SMC '06. IEEE International Conference on. Volume 4. (2006) 3534–3539
17. Hare, J.S., Lewis, P.H., Enser, P.G.B., Sandom, C.J.: A linear-algebraic technique with an application in semantic image retrieval. In Sundaram, H., Naphade, M.R., Smith, J.R., Rui, Y., eds.: CIVR. Volume 4071 of Lecture Notes in Computer Science., Springer (2006) 31–40
18. Bontcheva, K., Derczynski, L., Funk, A., Greenwood, M.A., Maynard, D., Aswani, N.: TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing, Association for Computational Linguistics (2013)
19. Maynard, D., Bontcheva, K., Rout, D.: Challenges in developing opinion mining tools for social media. In: Proceedings of @NLP can u tag #usergeneratedcontent?! Workshop at LREC 2012, Turkey (2012)
20. Miller, G.A., Beckwith, R., Felbaum, C., Gross, D., Miller, C.Miller, G.A., Beckwith, R., Felbaum, C., Gross, D., Miller, C.Minsky, M.: Five papers on WordNet. (1990)
21. Kalantidis, Y., Pueyo, L.G., Trevisiol, M., van Zwol, R., Avrithis, Y.: Scalable triangulation-based logo recognition. In: Proceedings of the 1st ACM International Conference on Multimedia Retrieval. ICMR '11, New York, NY, USA, ACM (2011) 20:1–20:7

22. Psyllos, A., Anagnostopoulos, C.N., Kayafas, E.: M-sift: A new method for vehicle logo recognition. In: Vehicular Electronics and Safety (ICVES), 2012 IEEE International Conference on. (2012) 261–266
23. Viola, P., Jones, M.: Robust real-time object detection. In: International Journal of Computer Vision. (2001)
24. Fasel, B., Luettin, J.: Automatic facial expression analysis: a survey. Pattern Recognition **36** (2003) 259 – 275
25. Pantic, M., Sebe, N., Cohn, J.F., Huang, T.: Affective multimodal human-computer interaction. In: Proceedings of the 13th annual ACM international conference on Multimedia. MULTIMEDIA '05, New York, NY, USA, ACM (2005) 669–676
26. Tian, Y.l., Kanade, T., Cohn, J.F.: Facial expression analysis. Handbook of Face Recognition (2005) 247–275
27. Du, S., Tao, Y., Martinez, A.M.: Compound facial expressions of emotion. Proc Natl Acad Sci U S A (2014)
28. Ekman, P., Friesen, W.: Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Palo Alto (1978)
29. Friesen, W., Ekman, P.: EMFACS-7: Emotional Facial Action Coding System. Unpublished manual; University of California, California (1983)
30. Ekman, P., Irwin, W., Rosenberg, E.R., Hager, J.C.: FACS Affect Interpretation Database (FACSAID). http://face-and-emotion.com/dataface/facsaid/description.jsp (1997)
31. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models – their training and application. Comput. Vis. Image Underst. **61** (1995) 38–59
32. Cootes, T., Edwards, G., Taylor, C.: Active appearance models. Pattern Analysis and Machine Intelligence, IEEE Transactions on **23** (2001) 681–685
33. Saragih, J.M., Lucey, S., Cohn, J.: Face alignment through subspace constrained mean-shifts. In: International Conference of Computer Vision (ICCV). (2009)
34. Lucey, P., Cohn, J., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on. (2010) 94–101
35. Chew, S., Lucey, P., Lucey, S., Saragih, J., Cohn, J., Sridharan, S.: Person-independent facial expression detection using constrained local models. In: Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on. (2011) 915–920
36. Ryan, A., Cohn, J.F., Lucey, S., Saragih, J., Lucey, P., De la Torre, F., Ross, A.: Automated facial expression recognition system. In: IEEE International Carnahan Conference on Security Technology. (2009)
37. Dhall, A., Goecke, R., Lucey, S., Gedeon, T.: Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In: Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on. (2011) 2106–2112
38. Parkhi, O., Vedaldi, A., Zisserman, A.: On-the-fly specific person retrieval. In: Image Analysis for Multimedia Interactive Services (WIAMIS), 2012 13th International Workshop on. (2012) 1–4
39. Dong, W., Wang, Z., Charikar, M., Li, K.: High-confidence near-duplicate image detection. In: ACM ICMR'12, ACM (2012) 1:1–1:8
40. Hare, J., Samangooei, S., Dupplaw, D., Lewis, P.: Imageterrier: An extensible platform for scalable high-performance image retrieval. In: ICMR 2012. (2012)
41. Hare, J., Samangooei, S., Dupplaw, D., Lewis, P.H.: Twitter's visual pulse. In: 3rd ACM International conference on multimedia retrieval. (2013) 297–298
42. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV **60** (2004) 91–110
43. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: ICCV. (2003) 1470–1477
44. Dong, W., Charikar, M., Li, K.: Asymmetric distance estimation with sketches for similarity search in high-dimensional spaces. In: SIGIR'08, ACM (2008) 123–130

45. Zontone, P., Boato, G., Hare, J., Lewis, P., Siersdorfer, S., Minack, E.: Image and collateral text in support of auto-annotation and sentiment analysis. In: TextGraphs-5: Graph-based Methods for Natural Language Processing, The Association for Computational Linguistics (2010) 88–92
46. Siersdorfer, S., Hare, J., Minack, E., Deng, F.: Analyzing and predicting sentiment of images on the social web. In: ACM Multimedia 2010, ACM (2010) 715–718
47. Wang, W., He, Q.: A survey on emotional semantic image retrieval. In: Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on. (2008) 117–120
48. Esuli, A., Sebastiani, F.: SENTIWORDNET: A publicly available lexical resource for opinion mining. In: Proceedings of LREC 2006. (2006)
49. Zerr, S., Siersdorfer, S., Hare, J., Demidova, E.: Privacy-aware image classification and search. In: SIGIR'12, New York, NY, USA, ACM (2012) 35–44
50. Zerr, S., Siersdorfer, S., Hare, J.: Picalert!: a system for privacy-aware image classification and retrieval. In: 21st ACM Conference on Information and Knowledge Management (CIKM 2012). (2012)
51. Siersdorfer, S., Zerr, S., Pedro, J.S., Hare, J.: Nicepic!: A system for extracting attractive photos from flickr streams. In: ACM SIGIR 2014, ACM (2014)
52. Pedro, J.S., Siersdorfer, S.: Ranking and classifying attractiveness of photos in folksonomies. In: 18th International World Wide Web Conference. (2009) 771–771