Challenges in developing opinion mining tools for social media

Diana Maynard, Kalina Bontcheva, Dominic Rout

Department of Computer Science
University of Sheffield
Regent Court, Sheffield, S1 4DP, UK
diana@dcs.shef.ac.uk

Abstract

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. Text mining systems originally designed for more regular kinds of texts such as news articles may need to be adapted to deal with facebook posts, tweets etc. In this paper, we discuss a variety of issues related to opinion mining from social media, and the challenges they impose on a Natural Language Processing (NLP) system, along with two example applications we have developed in very different domains. In contrast with the majority of opinion mining work which uses machine learning techniques, we have developed a modular rule-based approach which performs shallow linguistic analysis and builds on a number of linguistic subcomponents to generate the final opinion polarity and score.

1. Introduction

In this new information age, where thoughts and opinions are shared so prolifically through online social networks, tools that can make sense of the content of these networks are paramount. In order to make best use of this information, we need to be able to distinguish what is important and interesting. There are obvious benefits to companies, governments and so on in understanding what the public think about their products and services, but it is also in the interests of large public knowledge institutions to be able to collect, retrieve and preserve all the information related to certain events and their development over time. The spread of information through social networks can also trigger a chain of reactions to such situations and events which ultimately lead to administrative, political and societal changes.

Social web analysis is all about the users who are actively engaged and generate content. This content is dynamic, rapidly changing to reflect the societal and sentimental fluctuations of the authors as well as the ever-changing use of language. The social networks are pools of a wide range of articulation methods, from simple "I like it" buttons to complete articles, their content representing the diversity of opinions of the public. The user activities on social networking sites are often triggered by specific events and related entities (e.g. sports events, celebrations, crises, news articles, persons, locations) and topics (e.g. global warming, financial crisis, swine flu). 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, providing content appraisal via the social web, while social media mining itself can enrich archives, moving towards structured preservation around semantic categories.

Within this work, we focus on the challenges in the development of opinion mining tools which, along with entity, topic and event recognition, form the cornerstone for social web analysis in this respect. We discuss a variety of issues related to the adaptation of opinion mining tools to social media, and the challenges they impose on a Natural Language Processing (NLP) system, along with two example applications we have developed in 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 broadcasting organizations based in Germany: Sudwestrundfunk and Deutsche Welle). The approach we have developed forms part of a set of tools for the archiving of community memories and the long-term preservation of (multilingual) Social Web content. Based around a number of use cases in various domains, ultimately we aim to answer questions such as:

- What are the opinions on crucial social events and on the key people involved?
- How are these opinions distributed in relation to demographic user data?
- How have these opinions evolved over time?
- Who are the opinion leaders?
- What is their impact and influence?

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 they do not contain much contextual information and assume much implicit knowledge. Ambiguity is a particular problem 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, tend to be less grammatical than longer posts, contain unorthodox capitalisation, 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.

Most opinion mining techniques make use of machine learning, but this is problematic in applications such as ours where a number of different domains, languages and text types are involved, because models have to be trained for each one, and large amounts of training data are required for good results. Typically, classifiers built using supervised methods, e.g. (Boiy et al., 2007), perform well on polarity detection tasks, but when used in new domains, their accuracy reduces disastrously (Aue and Gamon., 2005). While some work has focused on adapating ML methods to new domains (Blitzer et al., 2007), this only really focuses on the use of different keywords in similar kinds of text, e.g. product reviews about books vs. reviews about electronics. Our entity-centric approach, on the other hand, makes use of rule-based NLP techniques, but in contrast to more traditional NLP approaches involving full parsing, we use a much shallower but more focused approach based around entity and event recognition, which lends itself better to non-standard text.

In the following section, we discuss some related work in the field of opinion mining and more generally, in the field of text mining from social media. We then describe in Section 3 the approach we have adopted, and some ofthe challenges faced in Section 4. In Section 5 we discuss evaluation issues and give some preliminary results, and finish with an outlook to the future in Section 6.

2. Related Work

(Pang and Lee, 2008) present a wide-ranging and detailed review of traditional automatic sentiment detection techniques, including many sub-components, which we shall not repeat here. In general, sentiment detection techniques can be roughly divided into lexicon-based methods (Popescu and Etzioni, 2005; Scharl and Weichselbraun, 2008; Taboada et al., 2011) and machine-learning methods, e.g. (Boiy and Moens, 2009). Lexicon-based methods rely on a sentiment lexicon, a collection of known and precompiled sentiment terms. Machine learning approaches make use of syntactic and/or linguistic features (Pak and Paroubek, 2010b; Go et al., 2009), and hybrid approaches are very common, with sentiment lexicons playing a key role in the majority of methods, e.g. (Diakopoulos et al., 2010). For example, (Moghaddam and Popowich, 2010) 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 (Mullaly et al., 2010). Several evaluations have shown the usefulness of contextual information (Weichselbraun et al., 2010; Wilson et al., 2009), and have identified context words with a high impact on the polarity of ambiguous terms (Gindl et al., 2010). A further bottleneck is the time-consuming creation of these sentiment dictionaries, though solutions have been proposed in the form of crowdsourcing techniques[1].

Recently, techniques for opinion mining have begun to focus on social media, combined with a trend towards its application as a proactive rather than a reactive mechanism. Understanding public opinion can have important consequences for the prediction of future events. One of the most obvious applications of this is for stock market predictions: (Bollen and Mao, 2011) found that, contrary to the expectation that if the stock markets fell, then public mood would also become more negative, in fact a drop in public mood acts as a precursor to a fall in the stock market.

Almost all the work on opinion mining from Twitter has used machine learning techniques. (Pak and Paroubek, 2010b) 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 (Go et al., 2009), 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, something which machine learning techniques can – to some extent – bypass with sufficient training data. For example. the Stanford NER drops from 90.8% F1 to 45.88% when applied to a corpus of tweets (Liu et al., 2010). (Ritter et al., 2011) also demonstrate some of the difficulties in applying traditional POS tagging, chunking and Named Entity Recognition techniques to tweets, proposing a solution based on LabeledLDA (Ramage et al., 2009).

There also exists a plethora of commercial search-based tools for performing sentiment analysis of tweets. Generally, the user enters a search term and gets back all the positive and negative (and sometimes neutral) tweets that contain the term, along with some graphics such as pie charts or graphs. Typical basic tools are Twitter Sentiment[2], Twends[3] and Twitrratr[4]. Slightly more sophisticated tools such as SocialMention[5] allow search in a variety of social networks and produce other statistics such as percentages of Strength, Passion and Reach, while others allow the user to correct erroneous analyses. On the surface, many of these appear quite impressive, and have the advantage of being simple to use and providing an attractive display with copious information about trends. However, such tools mostly aim at finding public opinion about famous people, sports events, products, movies and so on, but do not lend themselves easily to more complex kinds of opinion or to more abstract kinds of searches. Furthermore, their analy-

[1] http://apps.facebook.com/sentiment-quiz
[2] http://twittersentiment.appspot.com/
[3] http://twendz.waggeneredstrom.com/
[4] http://twitrratr.com/
[5] http://socialmention.com/

sis tends to be fairly rudimentary, performance can be quite low, and many of them do not reveal the sources of their information or enable any kind of evaluation of their success: if they claim that 75% of tweets about Whitney Houston are positive, or that people on Facebook overwhelmingly believe that Greece should exit the eurozone, we have no proof as to how accurate this really is.

Our approach to opinion mining takes inspiration from a number of sources. It is most similar to the work of (Taboada et al., 2011) in terms of technique, but because we focus on social media, we need to employ some different strategies to deal with the linguistic issues imposed. For example, we incorporate detection of swear words, sarcasm, questions, conditional statements and so on, while our entity-centric approach focuses the opinions on specific topics and makes use of linguistic relations.

3. Opinion mining

We have developed a series of initial applications for opinion mining from social media using GATE (Cunningham et al., 2002), a freely available toolkit for language processing. Based on the work described in (Maynard and Funk, 2011), which focused on identification in tweets of sentiments about political parties, we have extended this to a more generic analysis of sentiment about any kind of entity or event mentioned, within two specific domains: the current Greek financial crisis and the Rock am Ring rock festival in Germany in 2010. In both cases, we perform first a basic sentiment analysis by associating a positive, negative or neutral sentiment to each relevant opinion target, together with a polarity score. In the current scenarios, this could be any entity or event which is pertinent to the domain and use case. In the Rock am Ring corpus, this might be the overall event, a band or a band's particular performance at the concert, or some sub-event such as a light show that occurred during the performance. In the Greek crisis corpus, this might be a politician, an organisation, or an event such as a general strike or a relevant meeting that took place.

3.1. Entity extraction

The opinion mining application first requires that the corpus be annotated with entities and events. For this we have also developed a series of applications in GATE. We use a modified version of ANNIE (Cunningham et al., 2002), the default Named Entity (NE) recognition system in GATE, to find mentions of Person, Location, Organization, Date, Time, Money and Percent (though we only use the first three of these as potential opinion targets – the other entity types are used as additional indicators and, in some cases, feature values, in the linguistic patterns for opinion mining. We include some extra subtypes of Organization such as Band (for the Rock am Ring domain) and Political Party (for the Greek crisis domain), and have relaxed some of the settings to deal with the incorrectness of the English, though this has important ramifications. Detecting NEs in tweets, in particular, is challenging and we are currently performing some separate experiments about this. Enabling gazetteer lists to match against lowercase versions of proper nouns, for example, entails much greater ambiguity with common nouns. For example, the month "May" would be matched with the verb "may" – and even though we can also use a version of the POS tagger specially trained to deal with case-insensitive text, this is by no means guaranteed to work accurately all the time.

In addition to named entities, we also acquire a set of domain-specific terms using TermRaider[6]. This considers all noun phrases (NPs) – as determined by linguistic processing tools in GATE – as candidate terms, and then ranks them in order of termhood according to three different scoring functions: (1) basic tf.idf (Buckley and Salton, 2009) (2) an augmented tf.idf which also takes into account the tf.idf score of any hyponyms of a candidate term, and (3) the Kyoto score based on (Bosma and Vossen, 2010), which takes into account the number of hyponyms of a candidate term occurring in the document. All are normalised to represent a value between 0 and 100. We have not yet formally evaluated the three methods, though this is part of our planned future work, and indeed, it is possible that this may differ for differing domains or text types. Two further restrictions are placed. First, a candidate term is not considered as an entity if it matches or is contained within an existing Named Entity. Second, we set a threshold score above which we consider a candidate term to be valid. This threshold is a parameter which can be manually changed at any time – currently it is set to an augmented score of 45, i.e. only terms with a score of 45 or greater will be annotated as an Entity and used as input for the opinion mining and other tools.

3.2. Event recognition

In addition to entities, we also identify events to be used as possible targets for the opinions, and as input for other processes such as topic extraction (which fall outside the scope of this paper). Events can be expressed by text elements such as verbal predicates and their arguments ("The committee dismissed the proposal"), noun phrases headed by nominalizations ("economic growth"), adjective-noun combinations ("governmental measure"; "public money") and event-referring nouns ("crisis", "cash injection").

The pattern-based method we adopt involves the recognition of entities and the relations between them in order to find domain-specific events and situations, and is described more fully in (Risse et al., 2011). Currently we use only the events recognised by the top-down template-based approach, which consists of identifying a number of important events in advance, based on analysis of the user needs and manual inspection of the corpora. The template slots are pre-defined, and the values are entities extracted from the text as described in Section 3.1. In a semi-closed domain, this approach is preferable over the bottom-up approach, because it generates much higher precision results, while recall is not affected as significantly as in an open domain scenario.

Work on the event recognition is still very much in progress, though preliminary experiments showed very high precision (98% on a corpus of 1474 extracted events in the Greek

[6] http://gate.ac.uk/projects/neon/termraider.html

crisis dataset). We have not yet applied the event recognition to our Twitter or German datasets, where we expect to get somewhat lower results; however, these will be highly dependent on the quality of the entities extracted. Actually, we expect the quality of the event recognition (assuming correct entity detection) to be affected less by the typical problems associated with social media than the quality of the opinion mining and entity recognition tools, because we use such a shallow approach.

3.3. Sentiment Analysis

The approach we take for sentiment analysis is a rule-based one which is quite similar to that used by (Taboada et al., 2011), focusing on building up a number of subcomponents which all have an effect on the score and polarity of a sentiment. The main body of the opinion mining application involves a set of JAPE grammars which create annotations on segments of text. JAPE is a Java-based pattern matching language used in GATE (Cunningham et al., 2000). 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 (Miller et al., 1980). These have a feature denoting their part of speech, and information about the original WordNet synset to which they belong. The lists have been modified and extended manually to improve their quality: some words and lists have been deleted (since we considered them irrelevant for our purpose) while others have been added.

Once sentiment-bearing words have been matched, an attempt is made to find a linguistic relation between an entity or event in the sentence or phrase, and one or more sentiment-bearing words, such as a sentiment-bearing adjective modifying an entity or in apposition with it, or a sentiment-bearing verb whose subject or direct object is an entity. If such a relation is found, a Sentiment annotation is created for that entity or event, with features denoting the polarity (positive or negative) and the polarity score. The initial score allocated is based on that of the gazetteer list entry of the relevant sentiment word(s). The concept behind the scoring (and final decision on sentiment polarity) is that the default score of a word can be altered by various contextual clues. For example, typically a negative word found in a linguistic association with it will reverse the polarity from positive to negative and vice versa. Similarly, if sarcasm is detected in the statement, the polarity is reversed (in the vast majority of cases, sarcasm is used in conjunction with a seemingly positive statement, to reflect a negative one, though this may not necessarily be true of other languages than English). Negative words are detected via our Verb Phrase Chunker (e.g. “didn’t”) and via a list of negative terms in a gazetteer (e.g. “not”, “never”.). Adverbs modifying a sentiment adjective usually have the effect of increasing its intensity, which is reflected by multiplying the intensity factor of the adverb (defined in a gazetteer list) by the existing score of the adjective. For example, if “brilliant” had a score of 0.4, and “absolutely” had an intensity factor of 2, then the score of “brilliant” would increase to 0.8 when found in the phrase “absolutely brilliant”. Currently, the intensity factors are defined manually, but some of these could also be generated automatically where they are morphologically derived from an adjective (e.g. we could use the sentiment score of the adjective “brilliant” defined in our adjective list to generate an intensity factor for the adverb “brilliantly”).

Swear words, on the other hand, have a slightly more complex role. These are particularly prolific on Twitter, especially in the Rock am Ring corpus and on topics such as politics and religion, where people tend to have very strong views. First, 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 swearwords as hashtags (which also often contain more swear words in the main text of the tweet). The following rules are then applied:

- Swear words that are nouns get treated in the same way as other sentiment-bearing words described above. For example, in the tweet "Ed Miliband the world’s first talking garden gnome #f***wit", the word "f***wit" is treated as a sentiment-bearing word found in association with the entity "Ed Milliband".

- Swear words that are adjectives or adverbs are treated in the same way as regular adverbs, increasing the strength of an existing sentiment word. For example, if "awesome" scores 0.25, "fricking awesome" might score 0.5.

- Finally, any sentences containing swear words that have not been previously annotated are awarded a Sentiment annotation on the whole sentence (rather than with respect to an entity or event). For example, "Imagine saying how accepting of religions you are one day and the next writting a blog about how f***ed religions are" has no sentiment-bearing words other than the swear word, so the whole sentence is just flagged as containing a swearing sentiment. In this case, it is not easy to establish whether the sentiment is positive or negative – in the absence of any other clues, we assume such sentences are negative if they contain swear words and no positive words.

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.

3.4. Multilingual issues

Another artefact of social media is that corpora consisting of blogs, forums, Facebook pages, Twitter collections and so on are often multilingual. In our Rock am Ring corpus, comments and tweets can be in either English or German, while in the Greek financial crisis corpus, they can be in English or Greek, but also sometimes in other languages such as French. We therefore employ a language identification tool to determine the language of each sentence. The tool we use is a GATE plugin for the TextCat language identifier[7], which is an implementation of the algorithm described in (Cavnar and Trenkle, 1994). Each sentence is annotated with the language represented, and the application in GATE then calls one of two further applications, for English and German respectively, for each sentence being processed. If other languages are detected, then the sentence is ignored by the application and is not further analysed.
Language identification in tweets is a particular problem, due to their short length (140 characters maximum) and the ubiquity of language-independent tokens (RT (retweet), hashtags, @mentions, numbers, URLs, emoticons). Often, once these are removed, a tweet would contain fewer than 4 or 5 words, some would even have no "proper" words left. For English and German, we are currently achieving best results with the multinominal Naive Bayes language identifier by (Lui and Baldwin, 2011).

3.5. Adapting the tools for German

The approach we follow for processing German is very similar to that for English, but makes use of some different (though equivalent) processing resources in GATE. We have adapted the English named entity and term recognition tools specifically for German, using different POS taggers and grammars, for example. We also use the SentiWS dictionary (Remus et al., 2010) as the basis for our sentiment gazetteer. Currently, we do not perform event recognition in German (though this will be developed at a later stage), so opinions relate only to entities or to entire sentences and tweets.

4. Challenges imposed by social media

In addition to the factors already discussed, social media imposes a number of further challenges on an opinion mining system.

4.1. Relevance

Even when a crawler is restricted to specific topics and correctly identifies relevant pages, this does not mean that every comment on such pages will also be relevant. This is a particular problem for social media, where discussions and comment threads can rapidly diverge into unrelated topics, as opposed to product reviews which rarely stray from the topic at hand. For example, in the Rock am Ring forum, we also found comments relating to a televsion program that had been shown directly after the Rock am Ring event. Similarly on Twitter, the topics in which a user is interested can be very diverse, so it makes little sense to characterise "interesting" tweets for all users with a single lexical model.

[7] http://www.let.rug.nl/vannoord/TextCat/

There are a number of ways in which we can attempt to deal with the relevance issue. First, we could try to train a classifier for tweets or comments which are relevant, e.g. we might want to disregard tweets if they contain certain terms. Second, we can make use of clustering in order to find opinionated sentences or segments related to certain topics, and disregard those which fall outside these topics. This is probably the most promising approach, especially since we already make use of topic clustering algorithms within the wider project, although it does risk that some relevant comments might be left out.

4.2. Target identification

One problem faced by many search-based approaches to sentiment analysis is that the topic of the retrieved document is not necessarily the object of the sentiment held therein. This is particularly true of the online sentiment analysers discussed in Section 2, which make no connection between the search keyword and the opinion mentioned in the tweet, so that in fact while the polarity of the opinion may be correct, the topic or target of the opinion may be something totally different. For example, the day after Whitney Houston's death, TwitterSentiment and similar sites all showed an overwhelming majority of tweets about Whitney Houston to be negative; however, almost all these tweets were negative only in that people were sad about her death, and not because they disliked her. So the tweets were displaying dislike of the situation, but not dislike of the person. One way in which we deal with this problem is by using an entity-centric approach, whereby we first identify the relevant entity and then look for opinions semantically related to this entity, rather than just trying to decide what the sentiment is without reference to a target, as many machine learning approaches take. We use linguistic relations in order to make associations between target and opinion (for example, a target may be linked to a verb expressing like or dislike as its direct object, as in "I like cheese", or the opinion may be expressed as an adjective modifying the target "the shocking death of Whitney"). There are a number of ways in which sentences containing sentiment but which have no obvious target-opinion link can be annotated. Currently, we simply identify the sentence as "sentiment-containing" but make no assumption about the target. Future work will investigate further techniques for assigning a topic in such cases.

4.3. Negation

The simpler bag-of-words sentiment classifiers have the weakness that they do not handle negation well; the difference between the phrases "not good" and "good" is somewhat ignored in a unigram model, though they carry completely different meanings. A possible solution is to incorporate longer range features such as higher order n-grams or dependency structures, which would help capture more complete, subtle patterns, such as in the sentence "Surprisingly, the build quality is well above par, considering the rest of the features." in which the term "surprisingly" should partially negate the positive overall sentiment (Pang and Lee, 2008). Another way to deal with negation, avoiding the need for dependency parsing, is to capture simple

patterns such as “isn’t helpful” or “not exciting” by inserting unigrams like “NOT-helpful” and “NOT-exciting” respectively (Das and Chen, 2001). This work-around was implemented for tweets by Pak and Paroubek (Pak and Paroubek, 2010a).
For a rule-based system such as ours, we believe that the approach adopted, similar to that of (Taboada et al., 2011), is sufficient to capture most aspects of negation: indeed, Taboada’s evaluation appears to support this.

4.4. Contextual information

Social media, and in particular tweets, typically assume a much higher level of contextual and world knowledge by the reader than more formal texts. This information can be very difficult to acquire automatically. For example, one tweet in the political dataset used in (Maynard and Funk, 2011) likened a politician to Voldemort, a fictional character from the Harry Potter series of books. While the character is sufficiently well known to have its own Wikipedia entry, assimilating the necessary information (that Voldemort is considered evil) is a step beyond current capabilities, and we may have to just accept that this kind of comment cannot be readily understood by auomatic means.
One advantage of tweets, in particular, is that they have a vast amount of metadata associated with them which can be useful, not just for opinion summarisation and aggregation over a large number of tweets, but also for disambiguation and for training purposes. Examples of this metadata include the date and time, the number of followers of the person tweeting, the person’s location and even their profile. For example, we may have information about that person’s political affiliation mentioned in their profile, which we can use to help decide if their tweet is sarcastic when they appear to be positive about a particular political figure. Because each person registered on Twitter has a unique ID, we can disambiguate between different people with the same name – something which can be problematic in other kinds of text.

4.5. Volatility over Time

Social media, especially Twitter, exhibits a very strong temporal dynamic. More specifically, opinions can change radically over time, from positive to negative and vice versa. Within another project, TrendMiner[8], we are studying two highly dynamic opinion- and trend-driven domains: investment decisions and tracking opinions on political issues and politicians over time, in multiple EU states and languages. Since there is also correlation between the two domains, joint models of political opinions and financial market opinions also need to be explored.
To address this problem, the different types of possible opinions are associated as ontological properties with the classes describing entities, facts and events, discovered through information extraction techniques similar to those described in this paper, and semantic annotation techniques similar to those in (Maynard and Greenwood, 2012) which aimed at managing the evolution of entities over time. The extracted opinions and sentiments are time-stamped and stored in a knowledge base, which is enriched continuously, as new content and opinions come in. A particularly challenging question is how to detect emerging new opinions, rather than adding the new information to an existing opinion for the given entity. Contradictions and changes also need to be captured and used to track trends over time, in particular through opinion merging, which we turn to next.

[8] http://www.trendminer-project.eu

4.6. Opinion Aggregation and Summarisation

Another novel aspect to our work concerns the type of aggregation that can be applied to opinions to be extracted from various sources and co-referred. In classical information extraction, this can be applied to the extracted information in a straightforward way: data can be merged if there are no inconsistencies, e.g. on the properties of an entity. Opinions behave differently here, however: multiple opinions can be attached to an entity and need to be modelled separately, for which we advocate populating a knowledge base. An important question is whether one should just store the mean of opinions detected within a specific interval of time (as current opinion visualisation methods do), or if more detailed approaches are preferable, such as modelling the sources and strength of conflicting opinions and how they change over time. Effectively, we advocate here a form of opinion-based summarisation, e.g. displaying positive/negative opinion timelines, coupled with opinion holders and key features.
A second important question in this context involves finding clusterings of the opinions expressed in social media, according to influential groups, demographics and geographical and social cliques. Consequently, the social, graph-based nature of the interactions requires new methods for opinion aggregation.

5. Evaluation

Evaluation of opinion mining can be tricky, for a number of reasons. First, opinions are often subjective, and it is not always clear what was intended by the author. For example, we cannot necessarily tell if a comment such as "I love Baroness Warsi", in the absence of further context, expresses a genuine positive sentiment or is being used sarcastically. Inter-annotator agreement performed on manually annotated data therefore tends to be low, which affects the reliability of any gold standard data produced. While Amazon Mechanical Turk has been used for producing such gold standard annotated corpora, similar problems apply with respect to inter-annotator agreement, even if multiple annotations are produced for each document. Second, it is very hard to evaluate polarity scores such as the ones we produce: for example, we cannot really say how correct the score of 0.6012 awarded to a comment in the Rock am Ring forum about the band "In Flames" being the person’s favourite band is, or whether a score of 0.463 would be better. However, while these scores technically represent strength of opinion, we can view them instead as an indicator of confidence. So we would therefore expect the sentiments expressed with high polarity scores to have higher accuracy, and can tailor our evaluation accordingly, looking for higher accuracy rates as the polarity score increases.
As mentioned in Section 4, much of the success of an entity-centric opinion mining tool depends on the quality

of the entities and events extracted. Because we adopt a high precision strategy, at the potential expense of recall, we aim to minimise this effect. Because we risk missing some opinions, we also have a backoff strategy of identifying opinionated sentences which do not specifically map to an extracted entity or event. These give us some extra opinions, but risk being irrelevant or outside the scope of our interest.

We have not yet formally evaluated the opinion mining tools, other than for the political tweets dataset, whose results are reported in (Maynard and Funk, 2011). However, initial results look promising. We manually annotated a small corpus of 20 facebook posts (in English) about the Greek financial crisis (automatically selected according to certain criteria by our crawler) with sentiment-containing sentences, and compared these with our system generated sentiment annotations. Our system correctly identified sentiment-containing sentences with 86% Precision and 71% Recall, and of these correctly identified sentences, the accuracy of the polarity (positive or negative) was 66%. While the accuracy score is not that high, we are satisfied at this stage because some of the components are not fully complete – for example, the negation and sarcasm components still require more work. Also, this accuracy score takes into account both incorrect and correct sentiment-bearing sentences, since the two tasks are not performed independently (i.e. we are not assuming perfect sentiment sentence recognition before we classify the polarity of them). On the other hand, the named entity recognition is very accurate on these texts - our evaluation showed 92% Precision and 69% Recall. Since we aim for high Precision at the potential expense of Recall, and since we have further plans for improving the recall, this is most promising. Clearly, further and more detailed evaluation is still necessary.

6. Prospects and future work

While the development of the opinion mining tools described here is very much work in progress, initial results are promising and we are confident that the backoff strategies inherent in the incremental methodology will enable a successful system. We advocate the use of quite shallow techniques for much of the linguistic processing, using chunking rather than full parsing, for instance. While we could incorporate the Stanford parser to give us relational information, previous experience shows that the performance of such tools is dramatically reduced when used with degraded texts such as tweets. Furthermore, our methodology enables the system to be easily tailored to new tasks, domains and languages. On the other hand, the linguistic sub-components can also be used as initial pre-processing to provide features for machine learning, where such data is available, and we are currently experimenting with such techniques.

In previous work we have obtained good results using SVM-based machine learning (ML) from linguistic features for opinion classification (Funk et al., 2008; Saggion and Funk, 2009). We plan to experiment with similar data-driven techniques on tweets, although we would probably use the Perceptron algorithm instead, since it is faster and (in our experience) about as accurate for NLP. Our previous experiments were carried out on longer, somewhat more consistently edited texts (film, product and business reviews), which were quite unlike the highly abbreviated and inconsistent styles found in tweets. However, we obtained good results with unigrams of simple linguistic features, such as tokens and their lemmas, as well as with features derived from SentiWordNet values. With the additional features we already identify using our rule-based techniques, such as negative and conditional detection, use of swear words and sarcasm, we would expect to have some reasonable results. To carry out such experiments successfully on tweets, however, we would need a larger manually annotated corpus than the one previously used

As discussed earlier, there are many improvements which can be made to the opinion mining application in terms of using further linguistic and contextual clues: the development of the application described here is a first stage towards a more complete system, and also contextualises the work within a wider framework of social media monitoring which can lead to interesting new perspectives when combined with relevant research in related areas such as trust, archiving and digital libraries.

Acknowledgements

This work was supported by funding from the Engineering and Physical Sciences Research Council (grant EP/I004327/1) and the European Union under grant agreements No. 270239 (Arcomem[9]) and No. 287863 (TrendMiner[10]).

7. References

A. Aue and M. Gamon. 2005. Customizing sentiment classifiers to new domains: a case study. In *Proc. of the International Conference on Recent Advances in Natural Language Processing*, Borovetz, Bulgaria.

J. Blitzer, M. Dredze, and F. Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Annual Meeting-Association For Computational Linguistics*, page 440.

E. Boiy and M-F. Moens. 2009. A machine learning approach to sentiment analysis in multilingual web texts. *Information Retrieval*, 12(5):526–558.

E. Boiy, Pieter Hens, Koen Deschacht, and Marie-Francine Moens. 2007. Automatic sentiment analysis of on-line text. In *Proc. of the 11th International Conference on Electronic Publishing*, Vienna, Austria.

Johan Bollen and Huina Mao. 2011. Twitter mood as a stock market predictor. *IEEE Computer*, 44(10):91–94.

W. Bosma and P. Vossen. 2010. Bootstrapping language-neutral term extraction. In *7th Language Resources and Evaluation Conference (LREC)*, Valletta, Malta.

C. Buckley and G. Salton. 2009. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523.

W.B. Cavnar and J.M. Trenkle. 1994. N-gram-based text categorization. *Ann Arbor MI*, 48113:4001.

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

[10] http://www.trendminer-project.eu/

H. Cunningham, D. Maynard, and V. Tablan. 2000. JAPE: a Java Annotation Patterns Engine (Second Edition). Research Memorandum CS–00–10, Department of Computer Science, University of Sheffield, November.

H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*.

N. Diakopoulos, M. Naaman, and F. Kivran-Swaine. 2010. Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *IEEE Symp. on Visual Analytics Science and Technology (VAST)*, pages 115–122.

A. Funk, Y. Li, H. Saggion, K. Bontcheva, and C. Leibold. 2008. Opinion analysis for business intelligence applications. In *First Int. Workshop on Ontology-Supported Business Intelligence*, Karlsruhe, October. ACM.

S. Gindl, A. Weichselbraun, and A. Scharl. 2010. Cross-domain contextualisation of sentiment lexicons. In *Proceedings of 19th European Conference on Artificial Intelligence (ECAI-2010)*, pages 771–776.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*.

Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. 2010. Recognizing Named Entities in Tweets. *Science And Technology*, 2008.

M. Lui and T. Baldwin. 2011. Cross-domain feature selection for language identification. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 553–561, November.

D. Maynard and A. Funk. 2011. Automatic detection of political opinions in tweets. In Dieter Fensel Raúl García-Castro and Grigoris Antoniou, editors, *The Semantic Web: ESWC 2011 Selected Workshop Papers, Lecture Notes in Computer Science*. Springer.

D. Maynard and M. A. Greenwood. 2012. Large Scale Semantic Annotation, Indexing and Search at The National Archives. In *Proceedings of LREC 2012*, Turkey.

G. A. Miller, R. Beckwith, C. Felbaum, D. Gross, G. A. Miller, C.Miller, R. Beckwith, C. Felbaum, D. Gross, and M. Miller, C.Minsky. 1980. Five papers on WordNetk-lines: A theory of memory.

S. Moghaddam and F. Popowich. 2010. Opinion polarity identification through adjectives. *CoRR*, abs/1011.4623.

A.C. Mullaly, C.L. Gagné, T.L. Spalding, and K.A. Marchak. 2010. Examining ambiguous adjectives in adjective-noun phrases: Evidence for representation as a shared core-meaning. *The Mental Lexicon*, 5(1):87–114.

A. Pak and P. Paroubek. 2010a. Twitter as a corpus for sentiment analysis and opinion mining. *Proceedings of LREC 2010*.

A. Pak and P. Paroubek. 2010b. Twitter Based System: Using Twitter for Disambiguating Sentiment Ambiguous Adjectives. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 436–439. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Information Retrieval*, 2(1).

A. Popescu and O. Etzioni. 2005. Extracting product features and opinions from reviews. In *Proc. of the Conference on Empirical Methods for Natural Language Processing (EMNLP)*, pages 339–346, Vancouver, Canada.

D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. 2009. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA. Association for Computational Linguistics.

R. Remus, U. Quasthoff, and G. Heyer. 2010. SentiWS - A Publicly Available German-language Resource for Sentiment Analysis. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

T. Risse, S. Dietze, D. Maynard, and N. Tahmasebi. 2011. Using Events for Content Appraisal and Selection in Web Archives. In *Proceedings of DeRiVE 2011: Workshop in conjunction with the 10th International Semantic Web Conference 2011*, Bonn, Germany, October.

A. Ritter, S. Clark, Mausam, and O. Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proc. of Empirical Methods for Natural Language Processing (EMNLP)*, Edinburgh, UK.

H. Saggion and A. Funk. 2009. Extracting opinions and facts for business intelligence. *RNTI Journal*, E(17):119–146, November.

A. Scharl and A. Weichselbraun. 2008. An automated approach to investigating the online media coverage of US presidential elections. *Journal of Information Technology and Politics*, 5(1):121–132.

M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 1(September 2010):1–41.

A. Weichselbraun, S. Gindl, and A. Scharl. 2010. A context-dependent supervised learning approach to sentiment detection in large textual databases. *Journal of Information and Data Management*, 1(3):329–342.

T. Wilson, J. Wiebe, and P. Hoffmann. 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399–433.