Large-scale news entity sentiment analysis

Ralf Steinberger, Stefanie Hegele, Hristo Tanev & Leonida Della Rocca

European Commission – Joint Research Centre (JRC), Via E. Fermi 2749, 21027 Ispra, Italy
Ralf.Steinberger@ec.europa.eu

Abstract

We work on detecting positive or negative sentiment towards named entities in very large volumes of news articles. The aim is to monitor changes over time, as well as to work towards media bias detection by comparing differences across news sources and countries. With view to applying the same method to dozens of languages, we use linguistically light-weight methods: searching for positive and negative terms in bags of words around entity mentions (also considering negation). Evaluation results are good and better than a third-party baseline system, but precision is not sufficiently high to display the results publicly in our multilingual news analysis system Europe Media Monitor (EMM). In this paper, we focus on describing our effort to improve the English language results by avoiding the biggest sources of errors. We also present new work on using a syntactic parser to identify safe opinion recognition rules, such as predicative structures in which sentiment words directly refer to an entity. The precision of this method is good, but recall is very low.

1 Introduction

This work is being carried out in the context of our large-scale multilingual media monitoring effort EMM (Europe Media Monitor, Steinberger et al. 2017, 2009). EMM processes a daily average of 300,000 news articles in about 70 languages. EMM groups related articles, tracks them over time, categorises them, extracts information such as entities and reported speech by and about people, translates eleven languages into English and produces various types of statistics. The intention is to add graphs showing how positive or negative certain persons or organisations are being talked about, how this differs between news sources and countries, and how this changes over time. Due to the high number of languages covered, EMM works mostly with knowledge-poor methods, using language-agnostic rules. EMM therefore does not rely on part-of-speech tagging, syntactic parsers or large-scale dictionaries (Steinberger 2012). EMM uses machine learning for other purposes, but there is not sufficient suitable data to train our multilingual sentiment analysis components.

The practical requirement in the context of EMM is to achieve reasonable precision in positive and negative mentions. The strongly dominant class of neutral mentions in the news domain is less relevant and would at best serve to measure the overall heat around an entity (subjectivity level). Due to the large number of partially redundant news articles, recall is secondary. As a future goal, it would be good to also recognise sentiment separately for different aspects (Kumar & Sebastian 2012), such as types of policy of a politician. Table 1 lists snippets that show readers the types of phenomena we encounter in our work.

The term sentiment analysis is often used as the generic term to mean the computational treatment of opinion, sentiment and subjectivity in text (Pang & Lee 2008), but the terms are not identical: It is possible to state an opinion without expressing a sentiment (Soo-Min & Hovy 2004). We aim to detect polar (positive or negative) judgements or appraisals towards entities (our opinion target), independently of the opinion holder. We are mostly interested in cases where the journalists (representing the news source) express a biased judgement. However, as quoting other persons who make an evaluative statement is the journalist’s wilful choice, we currently make the simplifying approximation to equate the opinion holder with the author. By searching for polar words in the news text, we mostly aim to detect the newspaper’s framing bias; We do not currently consider epistemological bias (focusing on the believability of a proposition; Recasens et al. 2013). We have no means to detect content bias (not giving equal treatment to both sides in a political conflict; Entman 2007), such as the selection of stories a newspaper chooses to cover (tackled by Fortuna et al. 2009).
Sentiment analysis can be performed at different levels: It can be document, sentence, word and feature-based (Liu 2012). We only look at a window of up to six words around the entity in order to limit the potential impact of good or bad news (as opposed to sentiment) on the resulting sentiment value (see Section 3).

In the following sections, we first give an overview of related work (Section 2) and explain our method to perform sentiment analysis (3). We will present the evaluation results, including a detailed error analysis and the steps undertaken to reduce the error rate (4). Section 5 summarises additional experiments in which we learned and used syntactic patterns for sentiment analysis. In Section 6, we summarise and conclude.

2 Related work

Here, we want to mention some relevant work in the field of sentiment analysis, including annotation efforts and competitive events, especially in the news domain.

Sentiment analysis is usually based on lexical knowledge encoded explicitly as dictionaries and taxonomies or indirectly, as machine learning models. SentiWordNet (Esuli & Sebastiani 2006; Baccinella et al. 2010) is a lexical resource specific to sentiment analysis that assigns to each WordNet synset – sets of synonyms for groups of English words – three sentiment scores: positivity, negativity, or objectivity. Other approaches for sentiment lexicon building make use of bootstrapping (e.g. Banea et al. 2008). J Steinberger et al. (2012) used triangulation to generate polar dictionaries in fifteen languages.

Most researchers look at sentiment analysis as a text classification task (Pang & Lee 2008; Kiritchenko et al. 2014; Grimmer & Stewart 2013). Features used typically are word classes and word types such as adjectives, verbs and sentiment shifters. Other researchers like Vileares et.al. (2015) demonstrate the usefulness of using syntactic representations for carrying out sentiment analysis. Socher et al. (2013) present the so-called Recursive Neural Tensor Network, which – trained on annotated treebanks and considering sentiment compositionality – assigns polarity to whole sentences.

The dominant part of current academic sentiment analysis work applies supervised or semi-supervised classification techniques trained on manually annotated data such as movie or product reviews while there is relatively little work on the

| No. | Sentence with Entity (opinion target) and possible sentiment word |
|-----|---------------------------------------------------------------|
| 1   | Statement by the HR/VP Federica Mogherini on cessation of **hostilities** in Syria: |
| 2   | … and as **humiliated** leaders like Deng Xiaoping came back from their internal exiles … |
| 3   | …began in the 1960s, **inspired** by Chinese revolutionary leader Mao Zedong, and has cost thousands of lives. |
| 4   | … published a post that poked **fun** at the Microsoft web browser **Internet Explorer**. |
| 5   | **Donald Trump supports** the so-called “First Amendment Defense Act,” (FADA), a bill to enable Kim Davis-style **discrimination** against LGBTQ people nationwide. |
| 6   | **Chief Minister Shivraj Singh Chouhan** has **good** intentions but there’s little one man can do. |
| 7   | **Facebook** ended up hiring one of its **best** bug hunters, Reginaldo Silva … |
| 8   | The suit, filed in July 2012, alleged that **JPMorgan Chase** issued **false** and **misleading** statements regarding its trading activity … |
| 9   | **Margaret Thatcher’s former PR guru shocked** listeners when he told a **poor-taste** ‘joke’ … |
| 10  | There’s a simple reason why **Lewis Hamilton** is the **favorite** for this year’s Monaco Grand Prix, |
| 11  | It has taken out billboards in New York’s Times Square praising Charles Schumer […] for opposing the deal and **chastising** Senator Kirsten Gillibrand for backing it. |
| 12  | **General Hassan Firouzabadi** highlighted rifts between **Rouhani** and **critics** who say negotiations don’t serve Iran’s national interest … |
| 13  | In Ohio **Obama** was **not supported** by various groups of people. |
| 14  | The claims made by the three US carriers that Etihad Airways and other Gulf carriers are **damaging** their business … |
| 15  | **Oliver Letwin**, the Prime Minister’s **troubleshooter**, had advised **Cameron** to stay for the stability of the country. |

Table 1. Real-life news sentences. The task is to identify sentiment towards the entity.
news genre. Li et al. (2008) annotated 108 German political news documents involving German politicians following the annotation scheme developed in Wiebe et al. (2005). The Interest Group on German Sentiment Analysis provides 605 annotated sentences from a parliamentary corpus as training data for their 2016 IGGSAS Shared Task (Ruppenhofer et al. 2016). Deng & Wiebe’s (2015b) English language MPQA 3 corpus contains some 70 annotated texts, including news. The sentiment slot filling track at the NIST Text Analysis Conference (TAC) also contains a news part (Mitchell 2013). Unfortunately, resources are altogether relatively small. No annotated resources exist for the many languages we need to cover, and each annotated corpus is annotated with different objectives and definitions in mind. For instance, Li et al. (2008) use the words polarity, opinions and tones of voice, without defining what is deemed to be positive or negative. They list the word Krieg (war) as being negative and Erfolg (success) as positive. The word ‘success’ could indeed be seen as a positive judgement of the writer towards the entity, but in our view, the word ‘war’ usually refers to a factual event that has nothing to do with sentiment or judgement. See Section 3 for more detail on our viewpoint that good/bad sentiment should be separated from good/bad news content.

While a lot of sentiment-related work does not well define what positive or negative means, we feel that work around Soo-Min and Hovy (2004) and Wiebe et al. (2005) is more mature. Soo-Min and Hovy (2004) use a more conceptually motivated model. They describe an opinion as a “quadruple [Topic, Holder, Claim, Sentiment], in which the Holder believes a Claim about the Topic, and in many cases associates a Sentiment, such as good or bad, with the belief”. Sentiments always involve the Holder’s emotions or desires, and may be present explicitly (lexically) or only implicitly. Soo-Min & Hovy only try to identify those cases where the opinion holder is mentioned in the text while we are particularly interested in those cases where the author (the journalist) holds a biased or judgemental view. Wiebe et al. (2005, pp. 4f) aim to describe linguistic expressions of so-called private states, a term coined by Randolph Quirk that includes internal states that cannot be directly observed by others, such as opinions, emotions, sentiments, speculations and evaluations. They distinguish (a) explicit mentions of private states (e.g. the US fears a spill-over), (b) speech events expressing private states (Person A said “The report is full of absurdities”) and (c) expressive subjective elements. The latter are “used by people to express their frustration, anger, wonder, positive sentiment, mirth, etc., without explicitly stating that they are frustrated, angry, etc.”. Sarcasm and irony often involve expressive subjective elements. In their terms, experiencers hold attitudes toward targets (e.g. in John hates Mary). SenticNet (Cambria & Hussain 2015; Cambria et al. 2012) and WordNet-Affect (Strapparava & Valitutti 2004) provide more fine-grained information covering various aspects of emotion and sentiment. We can expect better multilingual resources for more informed sentiment analysis systems in the future, but producing such linguistics resources is much more time-consuming than producing lists of polar words.

Starting in 2013, there have been sentiment slot filling tracks as part of the NIST Text Analysis Conference (TAC) series. The 2013 task (Mitchell 2013) was to identify sentiment polarity together with opinion holder and opinion target in English language open domain texts, using news, forum texts and the general web. While this task differs from our own in that it focuses on cases where the opinion holder is explicitly mentioned in the text, it is interesting to see that the best system reached a precision of 10% and the best F-Score reached (by another system) was 13.15%. Performance was best in fora texts. In 2016, the track's objective was to identify beliefs or sentiments (private states) and the track was extended to include also smaller portions of Spanish and Chinese language texts. Due to the inclusion of beliefs, the task is even harder to compare to our own.

Deng & Wiebe (2015a) worked on an entity/event-level sentiment analysis system recognising and inferring both explicit and implicit sentiments. In addition to recognising polarity and opinion target, they also aim at recognising the opinion holder. They evaluate their system on the annotated English language MPQA 3.0 corpus (Deng & Wiebe 2015b). The best experimental results for this complex task yielded a precision of almost 38% for positive, and almost 36% for negative pairs. The best F-measure achieved was 24% and 33%, respectively.

3 Approach to news sentiment analysis

In a nutshell, our base sentiment analysis method searches for words from four different sentiment
classes in a bag-of-words window around the entities (For details, see J Steinberger et al. 2011). Subjectivity is assumed if a sentiment word is present in this window (see the discussion in Section 6). The added positive and negative sentiment values in that word window help identify the polarity regarding the mentioned entity. We distinguish the four classes positive/negative and highly positive/negative. Altogether, there are about 2,300 sentiment dictionary entries. These were created by combining in-house and various publicly available resources, by learning more terms using distributional semantics, and by then manually curating the resulting list. The best-performing combination of several existing sentiment dictionaries were empirically identified by using a training set (Balahur et al. 2010). Sentiment value calculation is compositional in our system as it also considers lists of about 125 intensifiers, diminishers and negators. The optimal window size (six words on both sides of the entity, but not passing sentence borders) was also empirically identified using a training set (Balahur et al. 2010). The cumulative sentiment values are normalised for the varying window size (in case of sentence borders). Due to a lack of large enough training data, we did not have the opportunity to further optimise the sentiment value assignment algorithm automatically nor the final lists of sentiment words or their relative weights. Some of the sentiment words contain wildcards in order to capture possible variants (e.g. intelligent).

An important issue for us was that good/bad news should not have an impact on the sentiment values for the entities that may be randomly associated with the event, e.g. when a politician visits an area affected by a natural disaster. We therefore kept factual words that may be perceived negatively such as war, death or injured out of the sentiment word lists and our algorithm ignores potential sentiment words that are part of the defining vocabulary of the article’s news category. For instance. When a news item was categorised into the news category NATURAL DISASTER, we ignore the category-defining words natural and disaster.

In order to evaluate the system performance, we produced a manually annotated news corpus consisting of 1274 sentences mentioning entities. Two annotators were asked to annotate the entity mentions as being positive (POS), negative (NEG) or neutral (NEUT), following the guidelines described in Balahur & Steinberger (2009). The annotators reached an agreement of 80%, resulting in a fair kappa agreement of 0.65 (see Table 2, second column, IAA). In comparison, the inter-annotator agreement for Portuguese news entity sentiment analysis before (Column 3) and after the improvement rules (Column 4, see Section 4), as well as by the Stanford tool (Base, Column 5). 2nd. baseline results: ALL NEUT, i.e. assigning only the neutral class.

![Table 2](image)

| Prec/Rec | IAA   | Original | Improved | Base/SSA |
|----------|-------|----------|----------|----------|
| ALL      | 0.80  | 0.737 / 0.737 | 0.742 / 0.742 | 0.760 / 0.760 |
| NEUT     | 0.86  | 0.788 / 0.800 | 0.788 / 0.910 | 0.788 / 0.788 |
| POS      | 0.78  | 0.436 / 0.323 | 0.431 / 0.294 | 0.431 / 0.294 |
| NEG      | 0.87  | 0.507 / 0.312 | 0.507 / 0.310 | 0.507 / 0.310 |
| POS/NEG  | 0.83  | 0.500 / 0.317 | 0.507 / 0.303 | 0.507 / 0.303 |
| Base/ALL NEUT | .72  | .72       | .72       | .72       |

Table 2. Inter-annotator agreement (IAA, column 2), Precision/Recall for English language entity sentiment analysis before (Column 3) and after the improvement rules (Column 4, see Section 4), as well as by the Stanford tool (Base, Column 5). 2nd. baseline results: ALL NEUT, i.e. assigning only the neutral class.

The third column of Table 2 (Original system) shows the system performance. The overall accuracy achieved was 74%. However, the precision for only the positive and the negative classes (POS/NEG) in the three-way classification was 50%, which is useless for practical purposes. The comparison in J Steinberger et al. (2011) showed very similar results for all seven languages tested. Column 5 (Base/SSA) shows a second baseline: the results of the Stanford Sentiment Annotator (SSA; Socher et al. 2013). SSA assigns polarity at the sentence level, using semantic compositionality, a sentence parser and the Recursive Neural Tensor Network, trained on a movie review sentiment treebank. The results in Column 5 were calculated by assigning the sentence polarity to the entities they contain. While the authors report 81% accuracy for movie reviews, SSA as we used it yielded 26% on our news corpus, and 19% for the polar cases POS/NEG. We observe that SSA classified an overwhelming 67% of sentences as being negative.
The difficulty of our task lies in the fact that only two out of nine possible combinations can be considered correct and that the neutral class (which is the only one we are not interested in) is by far the largest (72% of cases). Random assignment would yield 9.3% for only the polar cases POS and NEG. Considering the difficulty of the task and the odds at play, a precision of 50% is a good achievement, but it is not sufficient for practical purposes.

4 Error analysis and rules to improve the output

The entity opinion mining system described in Section 3 was run for more than one year (May 2015-April 2016), on altogether about 20 million English language news articles. The sentiment word forms most frequently found next to an entity are shown in Table 3. There are minor variations from one month to the next. The cumulated frequency for all 11,303 sentiment word forms is 11.76 million. The top ten most frequent sentiment word forms account for 12.3% of all sentiment terms found and the top one hundred for 42% of all sentiment word occurrences. As improving the sentiment value assignment accuracy of the most frequent terms should thus have a high impact on the overall performance of the system, we focused our effort on the ten most frequent sentiment words by first analysing the system output errors and then writing rules to avoid them. For this purpose, we used the SketchEngine (Kilgarriff et al. 2014) and HTML-based in-house tools to analyse the usage of each polar word.

4.1 Sentiment assignment errors and solutions implemented

We found the following major sources of errors:

1. **Wild cards capture wrong strings**: Our sentiment dictionaries contain a few words with wild cards to capture variants. To avoid words such as supporting-actor (support%) or intelligence-gathering (intelligen%), the wild cards in our sentiment dictionaries were refined and some wrong matches were included in a stop word list of terms that should not be used.

2. **Multi-word expressions and non-sentiment uses**: Some sentiment words also occur as multi-word expressions that do not have sentiment meaning, e.g. well in as well as. We used SketchEngine and the TenTen web corpus (included with the Sketch Engine service) to produce the most frequent word n-grams containing our top frequent sentiment words and selected those manually where we did not perceive the intended sentiment meaning of the word. We identified 52 such multi-word expressions (e.g. best known, with respect to, super PAC, artificial intelligence, well-known, etc.) and added them to a stop word list.

3. **Sentiment words in uppercase**: Uppercase words are usually part of entity names and should be ignored, e.g. crisis in the organisation name Judicial Crisis Network and super in Super Tuesday. To test the hypothesis, three evaluators not part of the team judged 40 sentences, of which 20 contained uppercase and 20 lowercase sentiment words. Their agreement with the system more than doubled when filtering out the uppercase words: Agreement rates with the system went from 20, 40 and 20% to 40, 75 and 55%, respectively. This confirmed our hypothesis.

| Word        | Freq.  | Perc.  |
|-------------|--------|--------|
| support     | 310446 | 2.68%  |
| won         | 183227 | 1.58%  |
| win         | 178328 | 1.54%  |
| well        | 165016 | 1.42%  |
| help        | 124610 | 1.07%  |
| best        | 101223 | 0.87%  |
| supporters  | 100810 | 0.87%  |
| accused     | 97665  | 0.84%  |
| victory     | 91059  | 0.78%  |
| agreed      | 87384  | 0.75%  |
| crisis      | 87295  | 0.75%  |
| warned      | 84082  | 0.72%  |
| good        | 82835  | 0.71%  |
| clear       | 75171  | 0.65%  |
| emergency   | 72486  | 0.62%  |
| better      | 70896  | 0.61%  |
| called for  | 67307  | 0.58%  |
| pressure    | 63878  | 0.55%  |
| lost        | 61996  | 0.53%  |
| cooperation | 59970  | 0.52%  |
| allies      | 59439  | 0.51%  |
| aid         | 57188  | 0.49%  |
| militants   | 57001  | 0.49%  |
| criticism   | 52584  | 0.45%  |
| winning     | 50878  | 0.44%  |
| wins        | 49233  | 0.42%  |
| failed      | 47805  | 0.41%  |
| criticized  | 46628  | 0.40%  |
| ally        | 45362  | 0.39%  |
| supporting  | 44996  | 0.39%  |

**Table 3.** Sentiment words and their occurrence frequency and percentage (out of all sentiment words) in 12 months of English language news analysis.
4. **Reported speech**: Sentiment words found inside direct or indirect speech must not be linked to the entity who is the author of the quotations because the entity is the opinion holder and not the opinion target. Markers of direct and indirect speech should thus act like sentence borders. Patterns to recognise indirect speech were added, consisting of reporting verbs combined with the conjunction *that*. This rule eliminates 1% of all entity mentions followed by a sentiment word. To test our hypothesis, three evaluators not part of the team judged 100 sentences containing sentiment words, of which 50 contained indirect speech. When filtering out the indirect speech sentences, their agreement with the system rose significantly, i.e. from 31, 40 and 67% to 54, 72 and 78%, respectively. This confirmed our hypothesis.

5. **Scope of sentiment words and passive constructions**: Some sentiment verbs have a scope that only extends in one direction. For instance, only the object (patient) of the verb *support* is positively affected by the action verb while the subject (agent) could be considered the ‘opinion holder’. In passive sentences, the roles are inverted (*A was supported by B*). We thus formulated filters ensuring that the system does not recognise a positive sentiment towards the agent. To test this hypothesis, three evaluators judged 95 sentences containing the verb ‘support’, of which 53 were then filtered out to avoid that the sentiment word was associated with the agent of the sentence. The agreement of the evaluators with the system output went up from 52, 44 and 51% to 57, 52 and 60%, respectively. Considering the success of this rule, we then used distributional semantics to identify lists of further verbs behaving like ‘support’. We checked the results manually and added them to the filtering rules (e.g. *criticise, accuse, praise, reject, blame*).

These were the main cases where the system regularly went wrong. However, there are numerous cases where polar words do not refer to the nearby entity and where we were not able to find a regularity that could be exploited in a filtering rule.

4.2 **Sentiment analysis results after incorporating these insights – Results**

After implementing the filters and improvements described in Subsection 4.1, we carried out a new evaluation to see whether the overall results have improved (See the fourth column in Table 2 - Improved). In the test set of 1274 sentences with entity mentions, there were 52 occurrences of the top ten sentiment words (the ones we focused our improvement efforts on), of which 19 were filtered out by the filtering rules 3 to 5 mentioned in Section 4.1. This filter improved the accuracy, but not significantly: precision for negative mentions improved by 1.35% to 59.2% and it diminished by 0.3% for positive mentions (43.1%).

Due to the large neutral portion of the test sentences, the sample of subjective sentences is too small. We therefore compiled an additional test set consisting of 200 potentially subjective sentences, i.e. sentences containing an entity name and at least one polar word in its word window. The system assigned 89 positive and 111 negative sentiment values to this set. Four evaluators annotated this collection. Two of them were part of the team and they had ample opportunities to come to a common understanding of the task, while the other two simply followed the annotation instructions. The agreement overlap between the six pairs of human annotators ranged from 67.5% to 81.5%, which is rather good even though the kappa scores only ranged from 0.13 to 0.40 because the neutral class was so dominant. However, the annotators agreed very little with the system output, namely in between 7% and 31% of cases. The in-house annotators, who had a clearer understanding of the task, agreed less with the system (7 and 13%) than the two outsiders (18 and 31%). Annotators agreed slightly more with our baseline, the Stanford Sentiment Annotator, i.e. 36, 36.5, 37 and 43.5%, but still clearly below the 50% threshold.

These results are extremely disappointing. They are worse than those of the previous evaluation carried out. It looks like the majority of entity mentions in the news sentence test set are neutral/objective even when they are mentioned in close vicinity to words from our sentiment dictionaries: Our four annotators marked only between 18% and 39% as either positive or negative in spite of the presence of sentiment words in the immediate vicinity of entities.

5 **Syntactic rules for news entity sentiment analysis**

The flaw of our system seems to be the recognition of subjective mentions. We furthermore observed that polar words close to the entity mentions do not necessarily refer to the entity itself, i.e. the bag-of-words approach is too coarse-grained. We therefore worked on identifying safe rules capturing syntactic constructions in which
there is a clear relationship between the polar word and the entity. We developed patterns for two predicative constructions (Rules 1 and 2) and two more generic syntactic paths (Rules 3 and 4), as shown in the following pattern list:

1. NAME is (who is, to be, was, may be...) SENTIMENT ENTITY is impressive.
2. SENTIMENT-adjective- NAME Impressive ENTITY.
3. SENTIMENT word and the Entity have a common verbal or nominal ancestor in the syntactic tree, e.g.

   ENTITY delivered an impressive speech
   (here ‘delivered’ is the common verbal ancestor of ‘impressed’ and ENTITY)

   The impressive speech of ENTITY
   (with ‘speech’ as a common nominal ancestor).
4. Sentiment word under the name in the syntactic dependency tree, e.g.

   intelligent politician ENTITY.

We analysed the test set of 200 sentences to evaluate our heuristics, using the Stanford dependency parser (Manning et al. 2014). We took as ground truth the two evaluators with the highest agreement. These were the two EMM team members, who had classified most sentences as neutral. They disagreed in only one of the sentences captured by the four rules, which they discussed and for which they then came to an agreement.

Altogether, the four rules matched 25 of the 200 sentences (12.5%). The precision was 44% and the recall 32%. Rules 1 and 2, which are the safest and most obvious, matched three times and all three analyses were correct (precision 100%) while Rules 3 and 4 matched 22 times, of which eight were correct (precision is 36%).

The imposed relationship between the sentiment word and the entity in Rules 3 and 4 is thus too vague to be useful. This shows that if the sentiment word does not directly syntactically refer to the entity, things can go wrong. Jiang et al. (2011) seem to have come to the same conclusion for Twitter sentiment analysis. More information would be needed to determine the correct sentiment judgement in such cases. This becomes obvious with the constructed example the intelligent/impressive liar ENTITY (matching Rule 4), where the generally positive words contribute to a negative judgement due to the negative noun liar.

6 Discussion and Conclusion

The EMM news entity sentiment annotation module was developed using linguistically shallow methods so that it could be applied to dozens of languages. The tool has an accuracy of 74%, which goes much beyond a baseline system trained for movie reviews (the Stanford Sentiment Annotator SSA, Socher et al. 2013). However, its precision is only 50% for positive and negative entity mentions (SSA: 19%; Random: 9.3%). In this paper, we presented a large-scale error analysis and hand-crafted rules to improve the results by avoiding the major errors. This did not significantly improve precision. Major challenges are the dominant number of neutral entity mentions (72%) and the fact that polar words often do not refer to the nearby entity. A separate subjectivity filter should be applied rather than assuming polarity in all cases where a sentiment word is present. We should also review our list of polar words to ensure that they have a polar sentiment value, rather than pointing to positive or negative news content (e.g. win, victory, support).

In separate experiments, we applied a syntactic parser and rules imposing a scope relationship between the polar word and the entity. While predicative structures (X is NEG, or the NEG X) yielded 100% precision, their recall is unacceptably small.

In view of sentiment analysis work trained on social media data such as movie reviews (Socher et al. 2013) reaching over 80% of accuracy, the results presented in this paper may look sub-standard. However, it does make a difference whether large numbers of manually classified reviews on a specific product are available to train the system or whether wide-coverage news is being analysed. Our comparative results reported in Table 2 (performance drop of the SSA tool from 80% on the training data to 26% on our data) show that tools cannot simply be ported to another domain.

Secondly, even our system achieves 74% accuracy when including the neutral class, but if the aim is to achieve a precision that goes significantly above 50% for the minority classes POS and NEG, the task is harder. The precision for various sub-tasks at evaluations and shared tasks (of course partially on more complex tasks) often does not go much above the 50% threshold either (Mitchell 2013; Ruppenhofer et al. 2016). The reported results furthermore typically include the neutral class. It looks like generic systems like ours are not yet ready to be deployed automatically for uncontrolled public display.
References

Baccinella Stefano, Andrea Esuli & Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. LREC’2010, pp. 2200-2204.

Balahur Alexandra & Ralf Steinberger. 2009. Rethinking Sentiment Analysis in the News: from Theory to Practice and back. Proceeding of WOMSA.

Balahur Alexandra, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik Van Der Goot, Matina Halkia, Bruno Pouliquen, Jenya Belyaeva. 2010. Sentiment analysis in the news. Proceedings of the 7th Intern. Conf. on Language Resources and Evaluation (LREC’2010), pp. 2216-2220. Valletta, Malta, 19-21 May

Banea Carmen, Rada Mihalcea & Janyce Wiebe. 2008. A bootstrapping method for building subjectivity lexicons for languages with scarce resources. Proceedings of LREC’2008.

Cambria Erik & Amir Hussain. 2015. SenticNet. In: Sentic computing. A common-sense-based framework for concept-level sentiment analysis, pp. 23-71. Springer

Cambria Erik, Andrew Livingstone & Amir Hussain. 2012. The hourglass of emotions. In: A. Esposito et al. (eds.) Cognitive Behavioural Systems. Springer LNCS 7403, pp. 144-157.

Deng Lingjia & Janyce Wiebe. 2015a. Joint Prediction for Entity/Event-Level Sentiment Analysis using Probabilistic Soft logic Models. Proceedings of EMNLP, pp. 179-189.

Deng Lingjia & Janyce Wiebe. 2015b. MPQA 3.0: An entity/event-level sentiment corpus. Proceedings of NAACL, pp. 1323-1328.

Domingos de Arruda Gabriel, Norton Trevisan Roman, Ana Maria Monteiro. 2015. An Annotated Corpus for Sentiment Analysis in Political News. Proceedings of Symposium in Information and Human Language Technology. Natal, RN, Brazil, pp. 101-110.

Entman Robert. 2007. Framing Bias: Media in the Distribution of Power. Journal of Communication 57 (2007) pp. 163-173.

Esuli A & Fabrizio Sebastiani. 2006. SentiWordNet: a high-coverage lexical resource for opinion mining. Technical Report. Institute of Information Science and Technologies of the Italian National Research Council.

Fortuna Blaž, Carolina Galleguillos & Nello Cristiani. 2009. Detection of bias in media outlets with statistical learning methods. In Ashok Srivastava & Mehram Sahami (eds.): Text Mining, Classification, Clustering, and Applications. Chapman and Hall, pp. 27-50.

Grimmer Justin & Brandon Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political Analysis 21(3), pp 267-297.

Jiang Long, Mo Yu, Ming Zhou, Xiaohua Liu & Tiejun Zhao. 2011. Target-dependent Twitter sentiment classification. Proceedings of ACL’2011, pp. 151-160.

Kilgarriff Adam, Vit Baisa, Jan Bušta, Miloš Jakubiček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlík, Vit Suchomel. 2014. The Sketch Engine: ten years on. Lexicography, Volume 1, Issue 1, pp 7–36.

Kiritchenko Svetlana, Xiaodan Zhu & Saif Mohammad. 2014. Sentiment analysis of short informal texts. Journal of Artificial Intelligence Research (JAIR) 50 (2014, pp. 723-762.

Kumar A. & T. M. Sebastian. 2012. Sentiment analysis: A perspective on its past, present and future. International Journal of Intelligent Systems and Applications, 4(10):1-14.

Liu B. 2012. Sentiment analysis and opinion mining. Morgan and Claypool Publishers: Synthesis Lectures on Human Language Technologies.

Manning Christopher, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard & David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60.

Mitchell Margaret. 2013. Overview of the TAC2013 Knowledge Base Population Evaluation: Sentiment Slot Filling. https://tac.nist.gov/publications/2013/.

Pang L. & L. Lee. 2008. Opinion mining and sentiment analysis. Foundations and trends in information retrieval 2 (1-2), 1-135.

Recasens Marta, Cristian Danescu-Niculescu-Mizil & Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. Proceedings of the ACL conference, pp. 1650-1659.

Socher Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher Manning, Andrew Ng & Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. Proceedings of EMNLP 2013, pp. 1631-1642. Seattle, Washington.

Soo-Min Kim & Eduard Hovy. 2004. Determining the sentiment of opinions. CoLing Conference, Geneva.

Steinberger Josef, Mohamed Elraham, Maud Ehrmann, Ali Hüriyetoglu, Mijail Kabadjov, Polina Lenkova, Ralf Steinberger, Hristo Tanev, Silvia Vázquez, Vanni Zavarella. 2012. Creating sentiment dictionaries via triangulation. Decision Support Systems 53(4), pp. 689-694.
Steinberger Josef, Polina Lenkova, Mijail Kabadjov, Ralf Steinberger & Erik van der Goot. 2011. Multilingual entity-centered sentiment analysis evaluated by parallel corpora. Proceedings of the RANLP Conference, pp. 770-775.

Steinberger Ralf, Bruno Pouliquen & Erik van der Goot. 2009. An Introduction to the Europe Media Monitor Family of Applications. In: Fredric Gey, Noriko Kando & Jussi Karlgren (eds.): Information Access in a Multilingual World - Proceedings of the SIGIR 2009 Workshop (SIGIR-CLIR'2009), pp. 1-8. Boston, USA.

Steinberger Ralf. 2012. A survey of methods to ease the development of highly multilingual text mining applications. Language Resources and Evaluation, Vol. 46 (2), pp. 155-176, Springer.

Steinberger Ralf, Martin Atkinson, Teófilo Garcia, Erik van der Goot, Jens Linge, Charles Macmillan, Hristo Tanev, Marco Verile & Gerhard Wagner. 2017. EMM: Supporting the analyst by turning multilingual text into structured data. In: Deggendorfer Forum zur digitalen Datenanalyse e.V. Verantwortung in der digitalen Datenanalyse, pp. 85-108. Berlin, Erich Schmidt Verlag.

Strapparava Carlo & Alessandro Valitutti. 2004. WordNet Affect: an affective extension of WordNet. LREC, vol. 4, pp. 1083-1086.

Vilares David, Miguel Alonso & Carlos Gómez-Rodríguez. 2015. On the usefulness of lexical and syntactic processing in polarity classification of Twitter messages. Journal of the Association for Information Science and Technology, 66: pp. 1799-1816.