Packed Feelings and Ordered Sentiments: Sentiment Parsing with Quasi-compositional Polarity Sequencing and Compression

Karo Moilanen[1] and **Stephen Pulman**[2] and **Yue Zhang**[3]

Abstract. Recent solutions proposed for sentence- and phrase-level sentiment analysis have reflected a variety of analytical and computational paradigms that include anything from naïve keyword spotting via machine learning to full-blown logical treatments, either in pure or hybrid forms. As all appear to succeed and fail in different aspects, it is far from evident which paradigm is the optimal one for the task. In this paper, we describe a quasi-compositional sentiment learning and parsing framework that is well-suited for exhaustive, uniform, and principled sentiment classification across words, phrases, and sentences. Using a hybrid approach, we model one fundamental logically defensible compositional sentiment process directly and use supervised learning to account for more complex forms of compositionality learnt from mere flat phrase- and sentence-level sentiment annotations. The proposed framework operates on quasi-compositional sentiment polarity sequences which succinctly capture the sentiment in syntactic constituents across different structural levels without any conventional *n*-gram features. The results obtained with the initial implementation are highly encouraging and highlight a few surprising observations pertaining to role of syntactic information and sense-level sentiment ambiguity.

1 INTRODUCTION

Language affords a wonderfully rich array of devices for expressing subjectivity, sentiments, affect, emotions, stances, opinions, arguments, points of view, perspectives, slurs, and the many other forms of non-factuality. From the viewpoint of a computational algorithm, non-factual content is bound to appear noticeably fuzzier than what is usually the case in traditional, more factual NLP tasks such as sentence breaking, part-of-speech tagging, or topic categorisation, to name a few. On the other hand, recent advances in computational Sentiment Analysis, Opinion Mining, and Affect/Emotion Analysis (and other related areas) have produced applications which, while still leaving much to be desired, are already highly useful in practice and can in some cases mimic human sentiment interpretation relatively well.

All proposals made in the above areas ultimately face the same fundamental challenge which is to determine what happens when individual expressions with rich (non-)sentiment properties interact with each other. A wide range of different solutions can be found encompassing mere frequency-based keyword spotting with no or naïve analytical additions, various machine learning approaches that have incorporated shallow-structural or -semantic features, and explicit direct fully- or shallow-compositional sentiment logics (§5).

If the goal is to be able to fully understand and account for the very behaviour of sentiment in language, then the task of explaining a given expression is to be approached using some form of principled logical reasoning that tries to *systematically* analyse all different parts of the expression in order to arrive at a logically defensible, coherent, and interpretable explanation in each case. Logical reasoning gives rise to a number of fundamental **compositional sentiment processes** many of which are simple enough to be modelled directly (e.g. [18], [13]). The most basic process involves **sentiment charge** which effectively involves inserting sentiment into an otherwise neutral expression. For example, when the neutral sentence *“[This report will make you* $\_\_\_{}_i$ *for hours]*$^{(N)}$*”* is modulated by a positive sentiment carrier (e.g. *“laugh*$_i^{(+)}$*”*), the **non-neutral propagation** process causes the latter to propagate its non-neutral sentiment across the entire sentence (vice versa for *“weep*$_i^{(\,-)}$*”*). Another obvious process is **null composition** which simply involves combining expressions displaying the same polarity (e.g. *“[evil]*$^{(\,-)}$*[wars]*$^{(\,-)}$*”* = *“[evil wars]*$^{(\,-)}$*”*). Somewhat less frequent is the **direct reversal** process in which reversive expressions reverse other expressions’ polarities (e.g. *“[avoid]*$^{[\neg]}$*[trouble]*$^{(\,-)}$*”* = *“[avoid trouble]*$^{(+)}$*”*). More challenging are the numerous cases where clashing non-neutral polarities interact: in these cases, some form of **conflict resolution** is necessary whereby the ensuing conflicts are resolved using either syntactic or semantic means (e.g. *“[benefit]*$^{(+)}$*[fraud]*$^{(\,-)}$*”* = *“[benefit fraud]*$^{(\,-)}$*”*).

Although it is still unclear which computational paradigm is optimal for practical purposes, explicit sentiment logics that implement the above kinds of fundamental sentiment processes have generally been observed to be very precise. They however commonly require manual rules which specify how individual expressions are to interact in the analysis, and are not unlikely to suffer from limited recall levels. Less focused machine learning approaches typically offer greater coverage but risk becoming too domain-dependent. They have yet to explain how the many contextual factors that ultimately govern the interaction of individual expressions across different structural levels are to be captured in a uniform and exhaustive manner.

In this paper, we seek to bridge these two paradigms and propose a hybrid sentiment learning and parsing framework for (sub)sentential sentiment analysis that implements only one of the above logical sentiment processes directly while leaving the rest to be learnt proba-

[1] Oxford University Computing Laboratory, UK. email: Karo.Moilanen@comlab.ox.ac.uk

[2] Oxford University Computing Laboratory, UK. email: Stephen.Pulman@comlab.ox.ac.uk

[3] University of Cambridge Computer Laboratory, UK. email: Yue.Zhang@cl.cam.ac.uk

bilistically from annotated sentiment data. The justification for a hybrid strategy is that the most basic compositional processes are simple enough to be modelled directly while the more complex ones may necessitate a more data-driven approach (cf. [4]). The framework is conceptually simple yet surprisingly powerful, and lends itself naturally to uniform sentiment parsing across words, phrases, and sentences. It is abstract enough to reduce domain and structural dependency effects but specific enough to capture one of the most important behavioural properties of sentiment. The proposed framework can be implemented easily without any complex linguistic processing as it requires only flat phrase- and/or sentence-level sentiment annotations, a sentiment lexicon, and, optionally, a part-of-speech tagger and a syntactic parser.

2 POLARITY SEQUENCE MODEL

Consider the following sample sentence (Ex. 1) (non-neutral and reversive words underlined).

(1) *["Our* lives$^{(+)}$*will* never$^{[\neg]}$*be the same again, having* lost$^{(-)}$*our* loved$^{(+)}$*ones and everything [we had] having been* destroyed$^{(-)}$*," Moussa told IRIN.]*$^{(-)}$

In a baseline approach, a classifier could first consider the polarity frequencies in the sentence (cf. [6]). It would however fail in this case since a POS/NEG tie count of 2 would ensue. In a more popular approach, the classifier could instead consider the distribution of various *n*-grams in the sentence: the unigram *"lost"* and the bigram *"and everything"* might, for example, help the classifier to make a decision if it has seen them amongst its negative polarity training examples (cf. [17]). Although they can reach moderate accuracy levels, the major drawback in *n*-gram models is that they seldom generalise well beyond the training data used, ignore all positional and temporal aspects of the sentiment carriers, and also perform markedly worse at lower (sub)sentential levels where evidence is always scarcer.

More complex features can then be harnessed to account for limited contexts around individual sentiment carriers by using crude, fixed windows (e.g. ±5 words) or by considering the sentiment around and above the sentence. For example, the fact can be exploited that the reversive adverb *"never$^{[\neg]}$"* affects the positive noun *"lives$^{(+)}$"* (via the head verb *"be"*). However, structural features still struggle to cope with the fact that sentiment carriers and other expressions modifying them can occur in any syntactic position, with many nested long-distance dependencies involved (cf. [25]). When more global structural features (e.g. checking if the sentence is surrounded by negative sentences or if it is in a negative document) are used (cf. [12]), evidence may erroneously be amassed from structures whose sentiment properties have nothing at all to do with each other. If positional information (e.g. the verb *"destroyed$^{(-)}$"* is the last sentiment carrier in the sentence) is included (cf. [15]), the problem remains that the most salient carrier can occur anywhere in the sentence. A further complication arises from the fact that many sentiment carriers suffer from context-dependent polarity ambiguity which confounds the problem even further.

2.1 Quasi-Compositional Sentiment Sequencing

On the basis of these kinds of complications that have hampered learning-based approaches, we instead investigate the possibility of an alternative, much simpler, route around the problem by dropping all but one of the conventional assumptions: specifically, we focus *solely* on the linear order in which atomic sentiment polarities occur in a sentence. If we represent the above sentence (Ex. 1) based on the prior out-of-context sentiment polarities of the words in it, the following **raw polarity sequence** representation emerges (Ex. 2):

(2) 1:NTR **2:POS** 3:NTR **4:REV** 5:NTR 6:NTR 7:NTR 8:NTR 9:NTR **10:NEG** 11:NTR **12:POS** 13:NTR 14:NTR 15:NTR 16:NTR 17:NTR 18:NTR 19:NTR **20:NEG** 21:NTR 22:NTR 23:NTR

Note that in addition to the three sentiment polarities proper (POS, NTR, NEG), the sentiment reversal potential (REV) of a word is used here as a fourth 'polarity' (cf. [4]). Raw polarity sequences such as this can then be turned into learning features by treating each step (i.e. **slice**) in the polarity sequence as a separate feature. However, because sentences and phrases vary a great deal in terms of their length (i.e. the number of raw feature slices that they yield), raw polarity sequences risk generating too sparse feature vectors and do as such necessitate very large amounts of training data to cover the probability of each of the four polarities occurring in each of the slices. Hence, we seek to employ some form of feature reduction instead.

The fundamental compositional sentiment process of **null composition** described in §1 offers a simple, yet logically defensible, means to shrink the feature space. If it is the case that two expressions which display the same polarity (e.g. *"[evil]$^{(-)}$[wars]$^{(-)}$"*) cannot but result in a compositional expression with the very same polarity (e.g. *"[evil wars]$^{(-)}$"*), then the same holds for three, four, and, by extension, *n* expressions. Hence, all subsequences in a raw polarity sequence that display the same consecutive polarity can axiomatically be collapsed into a single feature slice. We accordingly observe that the present sentence reduces into the following **compressed quasi-compositional polarity sequence** (Ex. 3) (with old and new slice IDs):

1:NTR **2:POS** 3:NTR **4:REV** 5:NTR **6:NEG** 7:NTR **8:POS** 9:NTR **10:NEG** 11:NTR

(3)

Polarity (sub)sequence		
Raw		Compressed
1:NTR	▷	1:NTR
2:POS	▷	**2:POS**
3:NTR	▷	3:NTR
4:REV	▷	**4:REV**
5:NTR 6:NTR 7:NTR 8:NTR 9:NTR	▷	5:NTR
10:NEG	▷	**6:NEG**
11:NTR	▷	7:NTR
12:POS	▷	**8:POS**
13:NTR 14:NTR 15:NTR 16:NTR 17:NTR 18:NTR 19:NTR	▷	9:NTR
20:NEG	▷	**10:NEG**
21:NTR 22:NTR 23:NTR	▷	11:NTR

From the raw polarity sequence that originally had 23 feature slices, a compressed quasi-compositional polarity sequence (i.e. *"Our*$_1$ *lives*$_2$ *will*$_3$ *never*$_4$ *be*$_5$ ... *lost*$_6$ *our*$_7$ *loved*$_8$ *ones*$_9$ ... *destroyed*$_{10}$ *Moussa*$_{11}$*"*) with only 11 feature slices (compression rate 52.17%) can therefore be derived. By 'quasi-compositional' we mean that the framework is aware of the fact that each compressed slice is composed of *n* sub-slices but does not attempt to analyse the composition: in other words, we jump directly from atomic prior sentiment (stemming from individual words) to more global sentiment without explaining the mapping(s) in between. The main assumption behind the quasi-compositional model is that, because of the null composition process, the compressed slices can still be expected to represent the *very same* sentiment information as their raw source slices: in the present example, although nearly half of the words were discarded, the sentiment information in the compressed 11 slices can be equated with that in the raw 23 slices.

Note that compressed polarity sequences can match a potentially very large number of unseen expressions regardless of which or how many words they contain because what is considered is the positions of the individual relevant sentiment polarities - *not the surface words* - in them. For example, a classifier trained on the present training example ought to be able to reason that an unseen chunk of text - be it a phrase with 11 words, a sentence with 25 words, or a document with 58 words - that contains compressed polarity slices ordered as `NTR_POS_NTR_REV_NTR_NEG_NTR_POS_NTR_NEG_NTR` can be negative. More importantly, if an unseen chunk of text fails to match any known sequence fully (e.g. when it is longer or shorter than any of the training examples), it is still likely to match many of the individual slice positions in the training data which means that the framework fails gracefully as the most optimal submatch can be expected in each case.

The advantage of the proposed polarity sequence model over simple *n*-gram modelling is that more information can in fact be captured because all key evidence can be accessed pertaining to the temporal (and hence positional) development of sentiment involving the smooth mixing, blending, figure/ground, and fading in/out behaviour amongst the three polarities ([10], [11]). Moreover, its advantage over more complex structural features is that polarity sequences may get rid of some unnecessary and untrue structural dependencies amongst words and syntactic constituents.

2.2 Feature Representation

In order for the compressed sequences to be used in supervised learning, we generate from each slice four separate features reflecting the polarity of the slice (i.e. `POS`, `NTR`, `NEG`, `REV`) represented with binary true/false values. The base polarity features can further be augmented with other information pertaining to various other properties of the words to which the feature slices point such as their word classes or grammatical roles. We consider further **non-sentiment-related features** from non-neutral and reversive words encompassing (i) word class tags (as output by a part-of-speech tagger) (§3.2), (ii) grammatical role tags (as output by a dependency parser) (§3.2), (iii) polarity word sense (WSD) ambiguity tags (as specified in a sentiment lexicon) (§3.1), and (iv) various combinations thereof. These additional non-sentiment-related features can be incorporated in two distinct ways. If **composite tags** are used, then additional non-sentiment-related evidence can be represented with more specific features. For example, the features from the above compressed polarity sequence can be enriched to include information such as the following (Ex. 4) (two sample slices shown):

	Word class	`2:POS:N|8:POS:ADJ|` ..
(4)	Syntax	`4:REV:ADV|6:NEG:MAIN-V|` ...
	WSD	`2:POS:NTRPOS|10:NEG:NONE|` ...

The classifier could then consider whether the eighth slice points to a positive adjective, whether the sixth one is a negative main verb, or whether the second slice points to a positive word that can also be neutral, for example. Another logical choice involves **parallel tags** amongst which additional non-sentiment-related evidence is scattered around multiple features. For example, parallel features such as the following can be had from the above compressed polarity sequence (Ex. 5) (two sample slices shown):

	Word class	`2:POS|2:N|8:POS|8:ADJ|` ...
(5)	Syntax	`2:POS|2:SUBJ|6:NEG|6:MAIN-V|` ...
	WSD	`2:POS|2:NTRPOS|10:NEG|10:NONE|` ...

In this case, the classifier could consider whether the second slice (i) is positive, (ii) points to a noun, (iii) functions as the subject in the sentence, and (iv) can be neutral or positive, respectively.

2.3 Training Data and Classifier

The learning models that we explore in this study were trained on two public domain data sets. The first ternary `POS`/`NTR`/`NEG` source, the *MPQA Opinion Corpus Version 2.0*[4] ([24]) (henceforth MPQA), yields 20822 (3993 (19.18%) `POS`, 7493 (35.99%) `NEG`, 9336 (44.84%) `NTR`) hand-labelled flat phrase- and sentence-level annotations from general news articles (inter-annotator agreement .72∼.82). Of the many different annotation types offered by the database, only *expressive subjectivity* and *direct subjectivity* annotations (*intensity* ∈ {*low*, *medium*, *high*, *extreme*}; *polarity* ∈ {*positive*, *negative*, *neutral*}) were included. Most of the training examples are short, with an average token count of ca. 2.69 (min. 1, max. 34, stdev. 2.29).

The second binary `POS`/`NEG` source, the *Sentence Polarity Data Set v1.0*[5] ([16]) (henceforth P&L), offers 10662 (5331 (50%) `POS`, 5331 (50%) `NEG`) flat sentence- and snippet-level annotations from (unverified) movie review star ratings mapped automatically onto binary sentiment polarities (inter-annotator agreement unknown). The P&L training examples are much longer, with an average token count of ca. 21.02 (min. 1, max. 59, stdev. 9.41).

In total, 18 models were trained from the two sources, in the conditions given in Table 1. The feature group label *pol* refers to base sentiment polarity features (§3.1), *wsd* to lexical polarity ambiguity features (§3.1), *pos* to word class features (§3.2), and *syn* to grammatical role features (§3.2). It can be seen that both training data sets could be captured with only a handful of slices (min. 20...27) which in turn translated into a small number of features (min. 58...99). Note that these figures are by a magnitude smaller than what would be the case if typical *n*-gram features were used as default unigrams would alone generate ca. 7800 (MPQA) vs. 18000 (P&L) features.

As a classifier of choice for the study, we used the Support Vector Machine implementation in the SVM.NET package with a linear kernel and all default parameters[6].

3 SENTIMENT PARSING

The previous sections illustrated the proposed framework that represents sentiment as compressed polarity sequences. The framework enables uniform sentiment parsing across words, phrases, and sentences without having to develop separate classifiers for different structural levels (e.g. running a sentence-level classifier to classify very short phrases). We combine the framework with a syntactic dependency parser to classify each individual syntactic constituent

[4] http://www.cs.pitt.edu/mpqa/

[5] http://www.cs.cornell.edu/People/pabo/movie-review-data/rt-polaritydata.tar.gz

[6] Johnson, M. (2008). SVM.NET 1.4. (www.matthewajohnson.org/software/svm.html). Based on Chang, C. & Lin, C. (2001). LIBSVM. (www.csie.ntu.edu.tw/~cjlin/libsvm/).

Table 1. Summary of learning models

Feature Groups	Feature Type	Features		Slices	
		MPQA	P&L	MPQA	P&L
pol	composite	58	99	20	27
pol.wsd	composite	183	394	22	33
pol.pos	composite	157	347	21	28
pol.syn	composite	380	977	26	32
pol.wsd	parallel	270	975	24	34
pol.pos	parallel	303	1048	33	47
pol.syn	parallel	501	1842	38	50
pol.wsd.pos.syn	composite	1031	3449	28	36
pol.wsd.pos.syn	parallel	1331	5507	55	78

in a piecemeal fashion, one sentence at a time. Fully compositional sentiment parsing can be achieved by allowing the sentiment polarity sequence model to base its decisions on its own previous decisions amongst constituents and their subconstituents in an incremental and recursive manner. We however focus in this initial study on the general properties of sentiment polarity sequencing at various non-interacting structural levels and leave the investigation of full composition for future work.

3.1 Sentiment Lexicon

The underlying sentiment knowledge that our framework draws on comes in the form of an extensive sentiment lexicon which contains 57103 manually classified entries tagged with various properties relevant to compositional sentiment interpretation across adjectives (22402, 39.2%), adverbs (6487, 11.4%), nouns (19004, 33.3%), and verbs (9210, 16.1%). Included are positive (21341, 37.4%), neutral (7036, 12.3%), and negative (28726, 50.3%) entries as well as reversive operators (1700, 3.0%) which are words and phrases that can directly reverse the polarity of a non-neutral expression (e.g. "*reduce*$^{[\neg]}$", "*no*$^{[\neg]}$", "*prevention*$^{[\neg]}$"). The lexicon also contains for each entry sentiment word sense ambiguity (WSD) tags that specify whether a given entry (i) unambiguously displays only one polarity across its senses (NONE) (e.g. "*woefully*$^{(\,-)}$"); is binary-ambiguous within the binary choice space (ii) positive or neutral (POSNTR) (e.g. "*brilliant*$^{(+)(N)}$"), (iii) negative or neutral (NEGNTR) (e.g. "*rat*$^{(N)(\,-)}$"), (iv) positive or negative (POSNEG) (e.g. "*proud*$^{(+)(\,-)}$"); or (v) is fully ternary-ambiguous (ANY) (e.g. "*high*$^{(N)(+)(\,-)}$"). The proposed framework is not tied to our current lexicon as any sentiment lexica can be used instead.

3.2 Grammatical Analysis

Each sentence is input into an initial grammatical analysis which involves part-of-speech tagging and syntactic dependency parsing. The chosen dependency parser[7] (i) tokenises the sentence into individual tokens, (ii) lemmatises them, (iii) assigns word class and other morphological features to them, (iv) creates syntactic links between them, and (v) labels the links according to their syntactic and dependency functions and types. The resultant raw dependency links between individual words in the sentence are converted into a flat, non-binary constituent tree in which each word in the sentence is treated as a head of a syntactic constituent for which sets of optional immediate (non-recursive) pre-head and post-head dependents are constructed. The proposed framework is not dependent in any way on this parser as any component that offers part-of-speech tags and marks syntactic constituent boundaries can be plugged in.

3.3 Recursive Sentiment Analysis

1st Pass. For each parsed sentence, we then assign prior sentiment polarities and polarity reversal values to all tokens based on the sentiment lexica (§3.1). All unknown words are asserted as neutral by default. Sentiment parsing involves first identifying plausible entry points into the dependency tree of the sentence which typically encompass (i) the main lexical head verb of the root clause, (ii) the head noun of a main clausal verbless NP, or (iii) a stranded word not linked to any other word in the sentence. The parser first descends recursively down to the lowermost atomic child leaf constituent under an entry constituent, and then climbs the tree upwards recursively to calculate a sentiment polarity for each intermediate constituent until all constituents - and hence the whole sentence - have been analysed.

When parsing a constituent, the parser follows a fixed head-dependents combination schema in combining the constituent head (H_i) with *k* pre- ($L_{i-k\,:\,i-1}$) and *j* post-head ($R_{i+1\,:\,i+j}$) dependents in a specific sequence, namely 1) first combining post-heads ([*R*]) with the head in a rightward direction (starting with the post-head nearest to the head), and 2) then combining the pre-heads ([*L*]) with the head-post-heads set ([*HR*]) in a leftward direction (starting with the pre-head nearest to the head). Each time a head is combined with a dependent, a chunk of text which reflects the surface words subsumed by the head-dependent pair is input into the sentiment sequence classifier. The resultant predicted polarity class label is then considered as the current global polarity in the analysis so far.

We accept the probabilistic predictions in all but one situation: in cases where a constituent head lacks any dependents (i.e. is made of just a singular word), we bypass the classifier and instead resort to the polarity assigned to the word in the lexicon. The reason for this simple exception is that there is no guarantee that the probabilistic classifier does not (i) override the prior polarity assigned to a word in the lexicon or (ii) render a neutral word non-neutral (e.g. inputting a NTR word into a binary POS/NEG model) in which case the framework would cease to be grounded on lexical knowledge. Note that our goal is to classify *combinations* of words, not individual words.

2nd Pass. The above 1st pass in the sentiment parsing process assigns sentiment to all syntactic constituents in a given sentence which ultimately results in all individual surface words displaying the final top-level compositional sentiment polarity/ies. In real-world use scenarios, the success (or the failure) of a sentiment algorithm will be judged based on whether or not the sentential polarities that individual surface words display make sense and 'read well'. It is unfortunately possible that some surface words end up displaying a polarity that appears incongruous with respect to the rest of the sentence. Such anomalies can stem from fragmentary grammatical analyses or arise when the classifier suggests a neutral polarity for a sentence even though it contains words which bear a *known* non-neutral polarity in the lexicon.

A further 2nd pass is therefore required to hide any traces of fragmentary or inconsistent analyses at the top sentence level. On the basis of the general tendency towards a coherent polarity flow within/across sentences (cf. [10], [11]), we accordingly account for 1) **neutral polarity gaps** (i.e. stranded neutral words amidst non-neutral words), and for 2) **non-neutral islands** (i.e. stranded non-neutral words that clearly disagree with the global majority sentiment of the sentence). For both gaps and islands, we simply execute a bidirectional lookup method around each incongruous surface word, and use the polarity evidence from their neutral/non-neutral neighbours as a heuristic masking polarity.

[7] Connexor Machinese Syntax 3.8.1. http://www.connexor.com/

4 EXPERIMENTS

Evaluating the performance of the proposed framework is not as straightforward as it seems. Firstly, because the sentiment sequence model is applied across all structural levels as part of exhaustive sentiment parsing, the targeted classification task is ultimately a ternary POS/NTR/NEG one for not all constituents are non-neutral: however, most public-domain gold standards come with binary POS/NEG annotations only. Accordingly, if a ternary classifier's output is evaluated against a binary gold standard (or vice versa), any conclusions that may be drawn are partial in the strictest sense. Secondly, since our framework assigns sentiment labels to all constituents in sentences, it is by no means clear which constituents ought to be evaluated. For example, if a gold standard contains expressions with arbitrarily chosen boundaries, there is no guarantee that the classifier's syntactic constituents map fully onto them (in fact they rarely do). As we are not aware of any manually-annotated and verified multi-level sentiment treebanks for English at the time of writing, we instead resort to three different gold standards which collectively shed light on the strengths and weaknesses of the framework at different structural levels. Due to these complications, we focus mainly on strictly binary evaluation conditions (whereby neither NTR predictions by the classifier nor NTR cases in the gold standard (if present) are considered) as they are much more indicative of core sentiment judgements.

4.1 Gold Standard Data Sets

Headlines [SEMEVAL]. The first data set comprises 1000 news headlines from the SemEval-2007 Task #14 annotated for polarity along the scale [-100...-1|0|1...100] (46.80% POS), 0.60% NTR, 52.60% NEG) (six annotators, inter-annotator agreement *r* .78) ([23])[8]. We included only the POS ([+1...+100]) and NEG ([-100...-1]) entries in the evaluation, and compare the classifier's sentential polarity against each headline. Ex. 6 illustrates sample headlines from the data set.

(6) [+32] *Test to predict breast cancer relapse is approved*
[-48] *Two Hussein allies are hanged, Iraqi official says*

Phrases [MPQA]. Evaluation targeting phrase-level expressions is based on the MPQA data set (§2.3) which we utilise for both ternary POS/NTR/NEG and binary POS/NEG evaluation. Ex. 7 illustrates a sample expression annotation in a sentence (annotation underlined).

(7) [LOW][POS] *Private organizations are also being encouraged to help fight sandstorms, according to the administration's vice-director Li Yucai.*

The MPQA expressions are considered in isolation without any contextual evidence from their hosting sentences in the MPQA database in order to avoid any subjective mappings or overlapping measures between the MPQA expression boundaries and our parser's constituents. In this condition, we compare the top-level polarity output by the classifier against each expression.

Snippets [P&L]. Further sentence- and snippet-level evaluation data come from the P&L data set (§2.3). Because a given snippet may consist of multiple sentences, we evaluate the majority 'document-level' polarity output by the classifier against each snippet in this condition. Ex. 8 illustrates a sample sentence from the data set.

(8) [NEG] *it wouldn't be my preferred way of spending 100 minutes or $7.00.*

4.2 Evaluation Measures and Baselines

A large number of different evaluation measures can be used to characterise the performance of the models, each of which highlights a different evaluative aspect. We hence evaluate the models using multiple complementary measures. The first measure family targets the conventional notion of 'accuracy' used in traditional factual classification tasks encompassing **Accuracy**, **Precision**, and **Recall** measures. For these, individual pairwise polarity decisions (POS vs. NOT-POS, NTR vs. NOT-NTR, NEG vs. NOT-NEG) were used. The second measure family focuses on different levels of **agreement** and **correlation** between human sentiment judgements and our models by calculating chance-corrected rates based on the standard **Kappa** *k*, **Pearson**'s *r* product moment correlation coefficient, and **Krippendorff**'s α reliability coefficient measures. In ternary POS/NTR/NEG classification, not all classification errors are equal because classifying a POS case as NTR is more tolerable than classifying it as NEG, for example. We lastly characterise three distinct **error types** between human *H* and algorithm *A*, namely 1) FATAL errors ($H^{(\alpha)}A^{(\neg\alpha)}$ $\alpha\in\{+\,\text{-}\}$), 2) GREEDY errors ($H^{(N)}A^{(\alpha)}$ $\alpha\in\{+\,\text{-}\}$), and 3) LAZY errors ($H^{(\alpha)}A^{(N)}$ $\alpha\in\{+\,\text{-}\}$).

The models are further compared against three baselines, namely **positive** (POS_BASE), **negative** (NEG_BASE), and **majority** sentiment using raw polarity frequency counting (FREQ_BASE).

4.3 Results

[SEMEVAL]. Starting with the short headlines, Table 2 highlights the performance of the models in the 2-way POS/NEG condition. In overall, the results are highly encouraging on both training data sets and are comparable with sample levels reported in other studies ([23])[9]. MPQA training data yielded clearly better scores than P&L data because (i) the former contains much more training data, and (ii) the MPQA expressions and the SEMEVAL headlines are of similar lengths. Both training data sets surpassed the POS_BASE (47.08) and NEG_BASE (52.92) baselines while the P&L models struggled to outperform the very high FREQ_BASE level at 71.53. Binary accuracy levels range from 71.03 to 77.94 while precision varies interestingly between the two polarities in that positive sentiment (72.47∼84.14) is more precise than negative sentiment (71.45∼76.94). Recall in turn displays a reverse pattern as positive sentiment has a considerably lower recall (62.80∼66.88) than negative sentiment (84.41∼92.41). Agreement levels point towards moderate levels at around 52.47∼54.49.

[MPQA]. Models trained on the P&L training data reached even more promising rates on the MPQA data set which is shown in Table 3 (2-way POS/NEG condition). All models surpassed the accuracy baselines (POS_BASE (34.76), NEG_BASE (65.24), FREQ_BASE (70.32)). The scores are especially significant because the slices from the MPQA and P&L training data differ considerably in length. Again, the models perform well against reported levels reached in other studies[10]. While binary accuracy rose to 84.73, agreement

[8] http://www.cs.unt.edu/~rada/affectivetext

[9] Cf. the highest reported 3-way figures in [23] are ∼55.10 (accuracy), ∼61.42 (precision), and ∼66.38 (recall). Note, however, that their evaluation conditions are not strictly identical with ours.

[10] Cf. the highest reported 2-way figures in [25]: 421 are ∼74.5 (POS precision), ∼87.8 (NEG precision), ∼77.8 (POS recall), and ∼98.3 (NEG recall). [4] report ∼88.5...90.7 2-way accuracies. Note, however, that their evaluation conditions are not strictly identical with ours.

Table 2. Experimental results on the SEMEVAL data set, 2-way POS/NEG condition (↑= boost over pol features)

	Trained on 3-way MPQA [20882], tested on 3-way Semeval headlines [784...841]								
Feature Groups	pol	pol wsd	pol pos	pol syn	pol wsd	pol pos	pol syn	pol wsd pos syn	pol wsd pos syn
Feature Types	composite				parallel			composite	parallel
Acc POS/NEG	**77.94**	74.88	76.74	77.30	76.68	76.96	77.38	73.20	74.79
Prec POS	**84.14**	82.09	80.50	79.46	76.17	82.13	81.74	76.29	76.75
Prec NEG	75.52	72.47	75.17	76.43↑	**76.94**↑	74.87	75.53↑	71.99	73.86
Rec POS	57.36	49.85	57.57↑	57.42↑	**62.80**↑	56.93	58.63↑	51.75	58.26↑
Rec NEG	**92.41**	92.39	90.21	90.30	86.34	91.19	90.70	88.52	86.98
Kappa	**52.24**	44.88	49.89	50.12	50.48	50.36	51.43	42.21	46.72
Pearson	**54.49**	48.01	51.57	51.64	51.08	52.38	53.15	44.09	47.85
Krippendorff	51.43	47.57	51.14	50.35	**52.76**↑	51.12	51.87↑	48.11	51.28

	Trained on 2-way P&L [10662], tested on 3-way Semeval headlines [994]								
Feature Groups	pol	pol wsd	pol pos	pol syn	pol wsd	pol pos	pol syn	pol wsd pos syn	pol wsd pos syn
Feature Types	composite				parallel			composite	parallel
Acc POS/NEG	70.52	**71.03**↑	65.90	69.11	67.40	70.62↑	70.82↑	62.47	64.29
Prec POS	69.40	72.06↑	72.01↑	70.18↑	68.95	70.66↑	**72.47**↑	68.06	63.75
Prec NEG	**71.45**	70.31	63.34	68.40	66.45	70.60	69.73	60.47	64.67
Rec POS	**66.88**	62.82	45.09	59.83	55.98	64.32	61.32	38.25	55.98
Rec NEG	73.76	78.33↑	**84.41**↑	77.38↑	77.57↑	76.24↑	79.28↑	84.03↑	71.67
Kappa	40.73	**41.44**↑	30.12	37.51	33.90	40.75↑	40.95↑	22.83	27.84
Pearson	40.75	**41.75**↑	32.29	37.89	34.46	40.90↑	41.40↑	25.21	28.03
Krippendorff	**52.47**	52.21	45.30	50.61	48.94	52.21	51.84	41.61	47.21

with human sentiment annotations is closer to substantial levels (57.84∼67.20). A clear asymmetry towards negative sentiment can be attested as both negative precision (90.60) and negative recall (89.20) are higher than positive precision (78.10) and recall (83.06) (cf. similar observations in [25]: 421).

[P&L]. The P&L data set interestingly appears more challenging for models trained on the MPQA training data as can be seen in the markedly lower levels shown in Table 4 (2-way POS/NEG condition). Binary accuracy decreased to 61.65 while agreement rates dropped to the level of only fair agreement (23.20∼43.50). Although they surpassed the POS_BASE and NEG_BASE baselines (50), the models are just below the FREQ_BASE baseline (61.57). A polarity asymmetry can once again be observed between higher positive precision (72.76 vs. 58.23 (NEG)) vs. higher negative recall (87.40 vs. 42.79 (POS)). The unexpectedly lower performance stems from the disparity in the number of slices in the (3-way) MPQA and (2-way) P&L data sets. An alternative conclusion drawable from the cross-training and -testing between the MPQA and P&L data sets is that the polarity sequence model may work better when the training data (P&L) contains more slices than the test data (MPQA). Note however that the P&L data set is replete with sarcasm, irony, and unknown words not found in our lexica.

Neutral Polarity. Since our goal was to maximise the amount of training data for the models, we employed the MPQA data set in its entirety. We moreover aimed at emulating real-world conditions by using strictly separate data sets for training and testing instead of cross-validation conditions of any kind (e.g. [4], [25]). Unfortunately, no unseen testing data with neutral polarity instances were then available for our experiments as only the MPQA data set contains ternary annotations. In order to estimate the neutral polarity performance of the models, we examined the relative performance of neutral polarity against non-neutral polarities using the base polarity *pol* model on the MPQA data set itself. Note that because we train and test on the same data set, the figures are understandably higher that what can be expected from unseen neutral annotations in the future. Nevertheless, many useful observations can be made based on the figures in Table 5. The inclusion of neutral polarity is likely to have an adverse effect on overall performance - an observation which concurs with the general trend in the area (e.g. [13], [25]). In our experiments, neutral recall was somewhat low (62.11) but its accuracy (72.03) and precision (71.72) were still high relative to the non-neutral levels. If we consider the error types in the ternary condition, only 14.14% of the errors were FATAL: the high level of GREEDY errors (52.15) indicates that the models may display oversensitivity towards non-neutral sentiment. For reference, we also report the corresponding ternary rates offered by the same model trained on binary P&L data. Note however that all neutral predictions in this condition come from singular words that bypassed the classifier altogether (see §3.3). The general pattern is the same, albeit somewhat more pronounced.

Features. We lastly consider the relative merits of individual feature groups across all data sets. The first clearly evident pattern is that mere polarity features (*pol*) are generally highly effective - especially considering that *no n-gram evidence was used in any form.* It is in fact surprising that so few features (58 (MPQA), 99 (P&L)) can even reach such high rates with highest accuracies touching on 84.73, precision levels up to 90.60, and recall levels up to 92.41 in some cases. More intriguing is the evidence pertaining to the expected utility of the extra non-sentiment-related feature groups. On the one hand, sentiment WSD, word class, and syntactic information do facilitate the analysis in many cases. On the other hand, they also hurt the performance of the base features in a number of cases. Although all of the extra features help in some condition, none of them can be said to help categorically. The single most useful supporting role is played by word-level sentiment WSD features which gave a boost most often in 24 conditions (13 composite, 11 parallel), indicating that the WSD tags can crudely mask the sentiment ambiguity amongst the slices. The support given by word class and syntactic information was not as high as expected since both boosted the base features in 19 conditions (word class: 10 composite, 9 parallel; syntax: 6 composite, 13 parallel). This in turn seems to suggest that either more training data are required or that morphosyntactic information is subservient to mere linear polarity sequences. Against the conventional

Table 3. Experimental results on the MPQA data set, 2-way POS/NEG condition (↑= boost over pol features)

	Trained on 2-way P&L [10662], tested on 3-way MPQA [10709]								
Feature Groups	pol	pol wsd	pol pos	pol syn	pol wsd	pol pos	pol syn	pol wsd pos syn	pol wsd pos syn
Feature Types	composite				parallel			composite	parallel
Acc POS/NEG	84.21	**84.73**↑	83.89	83.31	83.79	83.95	80.53	77.82	81.04
Prec POS	74.28	75.17↑	76.64↑	73.13	**78.10**↑	74.12	70.19	68.93	72.93
Prec NEG	90.38	**90.60**↑	87.68	89.63	86.51	89.98	86.59	82.06	85.16
Rec POS	82.76	**83.06**↑	76.49	81.41	73.48	81.95	75.43	64.69	71.39
Rec NEG	84.97	85.61↑	87.77↑	84.31	**89.20**↑	85.00↑	83.20	84.71	86.11↑
Kappa	65.94	**67.00**↑	64.29	64.00	63.57	65.31	57.61	50.13	57.79
Pearson	66.18	**67.20**↑	64.29	64.22	63.64	65.51	57.70	50.19	57.80
Krippendorff	57.58	**57.84**↑	55.59	56.89	54.56	57.26	54.25	50.04	53.14

Table 4. Experimental results on the P&L data set, 2-way POS/NEG condition (↑= boost over pol features)

	Trained on 3-way MPQA [20882], tested on 2-way P&L [9743...10313]								
Feature Groups	pol	pol wsd	pol pos	pol syn	pol wsd	pol pos	pol syn	pol wsd pos syn	pol wsd pos syn
Feature Types	composite				parallel			composite	parallel
Acc POS/NEG	60.37	60.13	61.01↑	59.86	**61.55**↑	60.12	60.42↑	59.00	60.67↑
Prec POS	**72.76**	72.48	69.08	69.91	68.93	68.46	68.22	67.76	67.00
Prec NEG	56.65	56.41	57.70↑	56.71↑	**58.23**↑	56.88↑	57.20↑	56.03	57.68↑
Rec POS	33.49	33.41	40.13↑	33.58↑	42.57↑	38.14↑	39.63↑	34.35↑	**42.79**↑
Rec NEG	**87.40**	87.17	81.97	85.76	80.67	82.28	81.38	83.65	78.72
Kappa	20.85	20.51	22.08↑	19.41	**23.20**↑	20.38	20.97↑	18.01	21.48↑
Pearson	24.78	24.38	24.33	22.69	**25.12**↑	22.75	23.11	20.70	23.04
Krippendorff	39.89	39.82	42.51↑	39.62	**43.50**↑	41.52↑	42.12↑	39.70	43.16↑

principle of crude force, wielding all of the features did not help as only 9 conditions benefited from them (2 composite, 7 parallel): in many cases, they too proved counterproductive. We hypothesise that this is due to the sparser feature spaces involved. Regarding which feature representation option - composite vs. parallel - is optimal, no firm conclusions can be drawn.

In overall, the boost given by the extra non-sentiment-related features over the base polarity features can range between only 1.02 (Pearson) and as much as 10.65 (negative recall) (cf. +1.18 (agreement), +1.58 (negative precision), +2.35 (Kappa), +3.61 (Krippendorff), +3.83 (positive precision), +9.31 (positive recall)). However, their adverse effects are much more pronounced, potentially ranging from as much as -5.76 (positive precision) to -28.63 (positive recall) (cf. -8.05 (agreement), -8.68 (negative recall), -10.86 (Krippendorff), -10.99 (negative precision), -15.99 (Pearson), -17.90 (Kappa)).

5 RELATED WORK

Sentence and Phrase-level Sentiment Analysis. A wide range of different approaches have been attempted. At the base level, mere frequency counting ([6]) with naïve analytical or learning additions ([3], [6], [10]) can offer moderate accuracies in some tasks. Various more complex machine learning approaches have incorporated shallow structural features ([1], [3], [25]), or joint classification models that target the structural co-dependency between individual sentences and documents using constrained inference ([12]). At the other end of the spectrum, a number of explicit direct fully- or shallow-compositional sentiment logics have been developed most of which rely on hand-written combinatory rules and lexical sentiment seeds in conjunction with semantic scope-driven valence shifters ([18]); fully compositional syntax-driven parsing ([13], [21]); structured inference-based learning with lexical, negator, and voting features ([4]); cascaded pattern matching with shallow phrasal chunking ([8]); learning-based topic classifiers with shallow phrasal chunking ([14]); verb-centric event frames with scored knowledge bases ([20]); or other heuristic linking and ranking patterns ([15]).

Positional Features. Even though they appear intuitively useful, positional features have so far been somewhat underrepresented in the area. Past attempts have focused on simple positional information within sentences ([9]), documents ([17]), or discourse ([22]). The solution closest to our sequence model is the sequential approach in [11] who model global document-level sentiment using a temporal trajectory function from local sentential polarities calculated by an Isotonic Conditional Random Field-based classifier. None of the above are driven by any compositional sentiment processes.

Feature Reduction and Compression. Various feature reduction techniques have been used in conjunction with sentiment learning. Typically, they operate on *n*-gram features and remove redundant or weak features through subsumption ([19]), abstraction ([7]), log likelihood ratio filters ([5]), or more sophisticated search criteria ([2]) amongst others. The guiding force behind our proposed feature reduction mechanism is in contrast the fundamental, linguistically justified, null composition principle. A conceptually analogous approach to sentiment compression is mentioned in [26] who, in measuring controversy in social media, construct polarity 'micro-state vectors' from words' polarity intensities and then similarly try to compress them. However, they leave all feature reduction decisions to standard compression algorithms agnostic of any compositional sentiment processes.

6 CONCLUSION

We have described a simple, yet effective, hybrid sentiment learning and parsing framework which is grounded on one basic logically defensible compositional sentiment process and which uses additional supervised learning to deal with more complex sentiment processes. The proposed framework, which offers a natural, yet principled basis for sentiment reasoning, operates on quasi-compositional sentiment polarity sequences which succinctly capture the sentiment in syntactic constituents across different structural levels without any conven-

Table 5. Experimental results on the MPQA data set, 3-way POS/NTR/NEG condition

	Tested on 3-way MPQA [20882]											
	Accuracy			Precision			Recall			Error Severity		
Trained on	POS	NTR	NEG	POS	NTR	NEG	POS	NTR	NEG	FATAL	GREEDY	LAZY
3-way MPQA [20882]	79.97	72.03	82.85	48.48	71.72	78.55	71.25	62.11	72.00	14.14	52.15	33.72
2-way P&L [10662]	77.37	65.51	71.89	44.73	79.05	57.96	76.36	31.39	79.66	19.06	72.19	8.76

tional *n*-gram features. It can be used for uniform sentiment classification across words, phrases, and sentences, and requires only simple flat phrase- or sentence-level sentiment annotations, a sentiment lexicon, and, optionally, a part-of-speech tagger and a syntactic parser. The results obtained with the initial implementation are highly encouraging and suggest that simple linear polarity sequence features alone operate effectively.

REFERENCES

[1] Apoorv Agarwal, Fadi Biadsy, and Kathleen Mckeown, 'Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams', in Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009), pp. 24–32, Athens, Greece, (March 30 - April 3 2009).

[2] Edoardo Airoldi, Xue Bai, and Rema Padman, 'Markov blankets and meta-heuristics search: Sentiment extraction from unstructured texts', in Advances in Web Mining and Web Usage Analysis: Revised Selected Papers from the 6th International Workshop on Knowledge Discovery on the Web (WebKDD 2004), 167–187, Seattle, WA, USA, (August 22-25 2006).

[3] Alina Andreevskaia and Sabine Bergler, 'CLaC and CLaC-NB: Knowledge-based corpus-based approaches to sentiment tagging', in Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pp. 117–120, Prague, Czech Republic, (June 23-24 2007).

[4] Yejin Choi and Claire Cardie, 'Learning with compositional semantics as structural inference for subsentential sentiment analysis', in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008), pp. 793–801, Honolulu, Hawaii, (October 25-27 2008).

[5] Michael Gamon, 'Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis', in Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004), pp. 841–847, Geneva, Switzerland, (August 23-27 2004).

[6] Minqing Hu and Bing Liu, 'Mining and summarizing customer reviews', in Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004), pp. 168–177, Seattle, Washington, USA, (August 22-25 2004).

[7] Mahesh Joshi and Carolyn Penstein-Rosé, 'Generalizing dependency features for opinion mining', in Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-AFNLP 2009), pp. 313–316, Singapore, (August 4 2009).

[8] Manfred Klenner, Angela Fahrni, and Stefanos Petrakis, 'Polart: A robust tool for sentiment analysis', in Proceedings of the 17th Nordic Conference on Computational Linguistics (NODALIDA 2009), pp. 235—238, Odense, Denmark, (May 14-16 2009).

[9] Lun-Wei Ku, I-Chien Liu, Chia-Ying Lee, Kuan hua Chen, and Hsin-Hsi Chen, 'Sentence-level opinion analysis by CopeOpi in NTCIR-7', in Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access (NTCIR 7), pp. 260–267, Tokyo, Japan, (December 16-19 2008).

[10] Hugo Liu, Henry Lieberman, and Ted Selker, 'A model of textual affect sensing using real-world knowledge', in Proceedings of the ACM International Conference on Intelligent User Interfaces (IUI 2003), pp. 125–132, Miami, Florida, USA, (January 12-15 2003).

[11] Yi Mao and Guy Lebanon, 'Isotonic conditional random fields and local sentiment flow', in Advances in Neural Information Processing Systems: Proceedings of the 2006 Conference (NIPS 2006), number 19, 961–968, Vancouver, British Columbia, Canada, (December 4-7 2007).

[12] Ryan McDonald, Kerry Hannan, Tyler Neylon, Mike Wells, and Jeff Reynar, 'Structured models for fine-to-coarse sentiment analysis', in Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007), pp. 432–439, Prague, Czech Republic, (June 23–30 2007).

[13] Karo Moilanen and Stephen Pulman, 'Sentiment composition', in Proceedings of the Recent Advances in Natural Language Processing International Conference (RANLP 2007), pp. 378–382, Borovets, Bulgaria, (September 27-29 2007).

[14] Kamal Nigam and Matthew Hurst, 'Towards a robust metric of opinion', in Computing Attitude and Affect in Text: Theory and Applications; Papers from the 2004 AAAI Spring Symposium (AAAI-EAAT 2004), Stanford, USA, (March 22-24 2004).

[15] Alexander Osherenko, 'Towards semantic affect sensing in sentences', in Proceedings of the Symposium on Affective Language in Human and Machine at the Communication, Interaction and Social Intelligence Convention (AISB 2008), pp. 41–44, Aberdeen, UK, (April 1-4 2008).

[16] Bo Pang and Lillian Lee, 'Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales', in Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 115–124, Ann Arbor, USA, (June 25-30 2005).

[17] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, 'Thumbs up? Sentiment classification using machine learning techniques', in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), pp. 79–86, Philadelphia, PA, USA, (July 6-7 2002).

[18] Livia Polanyi and Annie Zaenen, 'Contextual valence shifters', in Computing Attitude and Affect in Text: Theory and Applications; Papers from the 2004 AAAI Spring Symposium (AAAI-EAAT 2004), 106–111, Stanford, USA, (March 22-24 2004).

[19] Ellen Riloff, Siddharth Patwardhan, and Janyce Wiebe, 'Feature subsumption for opinion analysis', in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-06), pp. 440–448, Sydney, Australia, (July 22-23 2006).

[20] Mostafa Al Masum Shaikh, An analytical approach for affect sensing from text, Ph.D. dissertation, The Graduate School of Information Science and Technology, University of Tokyo, 2008.

[21] František Simančík and Mark Lee, 'A CCG-based system for valence shifting for sentiment analysis', in Advances in Computational Linguistics: Proceedings of 10th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2009), volume 41 of Research in Computing Science, 93–102, Mexico City, Mexico, (March 1-7 2009).

[22] Swapna Somasundaran, Janyce Wiebe, and Josef Ruppenhofer, 'Discourse level opinion interpretation', in Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008), pp. 801–808, Manchester, UK, (August 18-22 2008).

[23] Carlo Strapparava and Rada Mihalcea, 'Semeval-2007 task 14: Affective text', in Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pp. 70–74, Prague, Czech Republic, (June 2007).

[24] Janyce Wiebe, Theresa Wilson, and Claire Cardie, 'Annotating expressions of opinions and emotions in language', Language Resources and Evaluation, **39**(2-3), 165–210, (May 2005).

[25] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann, 'Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis', Computational Linguistics, **Volume 35**(3), 399–433, (September 2009).

[26] Zhu Zhang and Xin Li, 'Controversy is marketing: Mining sentiments in social media', in Proceedings of the 43rd Hawaii International Conference on System Sciences (HICSS-43 2010), pp. 1–10, Hawaii, USA, (January 5-8 2010).