A Comparative Study of Different Sentiment Lexica for Sentiment Analysis of Tweets

Canberk Özdemir and Sabine Bergler
CLaC Labs, Concordia University
1455 de Maisonneuve Blvd West
Montreal, Quebec, Canada, H3G 1M8
ozdemir.berkin@gmail.com, bergler@cse.concordia.ca

Abstract

We report on interoperability of different sentiment lexica with each other and with the linguistic notions *negation* and *modality* for sentiment analysis of tweets in a comprehensive ablation study and in competition results for SemEval 2015. Our approach performed well at the tweet level, but excelled in the presence of figurative language.

1 Introduction

Increasing interest in social media is reflected in SemEval competitions on sentiment analysis of tweets. Sentiment analysis categorizes text into positive or negative sentiment, possibly with an additional neutral category (Pang and Lee, 2008). Tweets use more informal and non-standard language than other text forms posing additional challenges. The winners of the past two years made heavy use of their specially designed NRC lexicon (Mohammad et al., 2013; Kiritchenko et al., 2014), a large lexical resource extracted from tweets with hashtags that are unmistakably positive or negative. This leads to the question we address here: is a bigger lexicon (proportionally) more useful? Is there something special about the NRC lexicon? Is a lexicon that is designed like the NRC lexicon but ten times its size more useful? And finally, can linguistic contexts *negation* and *modality* improve the lexicon and the final classification?

We compiled a NRC-inspired lexical resource, Gezi, of seven times the size of the NRC lexicon. We used several lexica in various combinations: Gezi, NRC, Bing Liu's lexicon (Hu and Liu, 2004), MPQA (Wilson et al., 2005), and aFinn (Nielsen, 2011), and add negation and modality sensitive features, performing comprehensive ablation experiments. The system competed in SemEval 2015 and ranked 9/40 in Task 10B, sentiment classification of tweets, and 1/35 in Task 11, tweets featuring figurative language.

2 Previous Work

Since Pang and Lees pioneering work on movie review classification into thumbs up—thumbs down (Pang et al., 2002), the major resource for sentiment determination was a sentiment lexicon, modelled after the independently and previously created Harvard General Inquirer (Stone et al., 1966), a list of words labelled as positive or negative sentiment carriers. Rule-based approaches yielded strong baselines that depended mainly on the coverage of the lexicon used, leading to various efforts to compile dedicated sentiment lexica (Esuli and Sebastiani, 2006; Wilson et al., 2005). The growing number of sentiment laden text on social media led to more efforts to annotate corpora, enabling machine learning approaches which monopolize the current exercises on sentiment annotation of tweets at SemEval (Rosenthal et al., 2015; Rosenthal et al., 2014; Nakov et al., 2013). The lexicon is still the major tool used, and for the non-standard use of language encountered in tweets, special resources have been compiled using the annotations displayed by the tweets themselves. Go et al. (2009), for instance, collected corpora using tweets containing positive or negative emoticons. In a similar way, Kiritchenko et al. (2014) use selected positive and negative hashtags to retrieve positive or negative tweets, computing association scores to the words occurring in tweets of each polarity. The resulting NRC lexicon was used by the winning team in SemEval 2013 and

Proceedings of Recent Advances in Natural Language Processing, pages 488–496,
Hissar, Bulgaria, Sep 7–9 2015.

2014, together with a simple negation feature.

The attention paid to sentiment in tweets led to the development of the CMU tagger Gimpel et al. (2011), a tokenizer and a POS tagger for tweets, as well as a named-entity recognizer for tweets (Ritter et al., 2011).

3 SemEval Datasets

The datasets for the SemEval exercises have been annotated using Amazon's Mechanical Turk[1] for Task 10 and CrowdFlower[2] for Task 11. The resulting annotations include, as expected, mislabelings and borderline judgements in the gold standard, such as:

labelled as negative *I haven't eaten chicken nuggets since I was like 6 or 7.. Who wants to get some McDonald's with me tomorrow?*

labelled as neutral *Class early in the mornjng =\it's bedtime! But do get to see my Sam tomorrow :)*

Polarity Classification Dataset Tweets with at least one term of SentiWordNet (Esuli and Sebastiani, 2006) association score greater than 0.3 or less than -0.3 form the corpus that is then manually labelled as positive, negative, or neutral.

The different test sets for 2013 (Nakov et al., 2013), 2014 (Rosenthal et al., 2014) and 2015 (Rosenthal et al., 2015) show a skewed distribution: with the exception of 2014, the majority of test cases are neutral and negative tweets form the smallest class, with the distribution changing slightly from year to year, see Table 1, where 'tw' stands for 'tweet', 'lj' for 'LiveJournal' entries, and 'sarc' for 'sarcastic tweets', different sources for test data for comparison.

Dataset	Positive	Negative	Neutral
tw-train	3,662	1,466	4,600
tw-dev	575	340	739
2013-tw-test	1,572	601	1,640
2013-sms-test	492	394	1,207
2014-tw-test	982	202	669
2014-sarc-test	33	40	13
2014-lj-test	427	304	411
2015-tw-test	1,038	365	987

Table 1: Dataset composition for Task10B.

Figurative Language Dataset The training set, consisting of 8000 tweets containing 5000 sarcastic, 1000 ironical and 2000 metaphorical tweets, was annotated on an 11 point scale (-5, . . . , +5) and released in two formats: tweets with integer or real-valued sentiment scores. The nature of figurative language tends to be negative Ghosh et al. (2015), Table 2 shows the distribution of instances for each integer sentiment score for training and test set.

Sentiment Value	Test Size	Training Size
-5	4	4
-4	99	434
-3	836	2,741
-2	1,540	2,546
-1	679	811
0	297	297
1	168	171
2	154	206
3	200	107
4	110	52
5	4	5

Table 2: Composition of datasets for Task 11.

4 Linguistic Notions

Following the successful use of a simple negation heuristic in (Kiritchenko et al., 2014), we further develop the use of the linguistic phenomena *negation* and *modality*. Negation and modality change the effect of the terms that occur in their scope, even though this change is not always one of total sentiment reversal for negation (a) or weakening for modality (b).

a. *Just watched the whole 2nd season of AHS in less then 24 hours. I'm not even ashamed.*

b. *Max might have to get put down tomorrow <3 absolutely heart breaking if I have to see my puppy go. Love you Maxy*

We use modality triggers *would like, would love, should, ought to, must, may, might, could, will, would, can, ca, cant, cannot, able, unable*; negation triggers from Rosenberg (2013); scope rules of Rosenberg et al. (2012).

Negation The simplest instances of negation parallel the logic operator: negation reverses the truth value of a proposition (*I'll do the dishes for you — NOT!*) but in natural language, usage is more varied, and negation is used to create contrast along other dimension, not only truth value, but also veridicity, and belief (*I don't believe that she did the dishes for you.*), to name but two. The degree to which a basic proposition is challenged is even more nuanced when modality (*I could do the dishes for you if you*

[1] https://www.mturk.com/mturk/
[2] http://www.crowdflower.com/

could take the garbage out.) and negation interact (*She might not have done the dishes for you.*). This impacts tasks of information extraction from on-line texts and while these phenomena have long been neglected as comparatively rare and benign in information retrieval contexts, precision-oriented information extraction has addressed negation and recently modality in a series of challenge tasks (BioNLP Shared Tasks 2008-2010 (Kim et al., 2009), CoNLL 2010 (Farkas et al., 2010), *Sem(Morante and Blanco, 2012), QA4MRE (Morante and Daelemans, 2012)).

Negation and modality affect sentiment (*I am not happy!* does not convey positive sentiment). Even simple negation heuristics are beneficial: consider the scope of the negation to span from a negation trigger to the next punctuation mark (Mohammad et al., 2013) or to occupy a fixed window around the negation trigger (Günther and Furrer, 2013). Our system uses a syntax-aware negation trigger and scope detection system developed by Rosenberg (2013).

The effect that negation is interpreted to have on the interpretation of a text varies. Kennedy and Inkpen (2005) encode negation as a simple reverser of polarity values (multiplying them by -1). However, negation does not always reverse the effects of the sentiment carriers, as the case of judgements illustrates: *This isn't awful.* does not mean *This is fantastic.* Since negated sentiment carriers do not default to one fixed resulting sentiment value but have to be assessed in their linguistic context, we do not resolve the negation numerically, but encode its occurrence in a separate feature (*negated-positive, negated-negative, negated-neutral*), a technique similar to Kennedy and Inkpen (2006).When computing the association scores for our Gezi lexical resource, a negation context results in multiplying sentiment association scores of sentiment carriers by -0.5, an empirically derived value.

Modality Modality indicates possibility, it dampens the asserted veridicity of a statement, often accompagnied by the reason for the hedging: second hand information, belief, hypothetical, ... In utility texts like newspaper articles or UNIX documentation, modality is a rare phenomenon. But in journal articles in the life sciences or in tweets, it is frequent and carries important meaning aspects. The BioNLP Shared Task series (Kim et al., 2009) paid special attention to speculative language, and QA4MRE (Morante and Daelemans, 2012) additionally addressed the interaction of negation and modality. Following Rosenberg et al. (2012), whose treatment at the QA4MRE pilot dominated the competition, we treat modality the same as we treat negation: a trigger list and scope annotation indicate the modalized material and we represent this and its interaction with negation by doubling our encoded features to include for example *mod-positive, negation-positive, mod-negation-positive* (see Table 5).

5 Lexical Resources

A number of sentiment lexica are available and have been used in various systems. To our knowledge, they have not been compared critically on the same task to assess their respective contribution alone or in combination. We perform such a comparative ablation exercise on some of the more widely used lexica in order to assess our own new lexical resource, Gezi.

5.1 Manually Compiled Lexica

We include MPQA lexicon and Bing Lius Opinion Lexicon, which includes MPQA entries and thus provides a first means to compare how size impacts performance. To complete the picture, we also use the much smaller aFinn lexicon.

MPQA (Wilson et al., 2005), manually compiled with prior polarities for over 8000 words, distinguishing *positive, negative*, and *neutral*. The terms also have pseudo-POS tag information for disambiguation purposes.

Opinion Lexicon of Bing Liu (Hu and Liu, 2004), manually selected lexicon of around 6800 terms, only *positive* and *negative*.

aFinn (Nielsen, 2011), lexicon of words manually rated for valence scores with an integer between -5 and 5 together with their prior polarities, around 2500 words.

5.2 Automatically Compiled Lexica

NRC Hashtag Sentiment Lexicon (Mohammad et al., 2013) This open source lexicon was key in the winning entry for the last two years. It is a large, automatically compiled resource that uses seed hashtags that carry unambiguous, strong sentiment as proxy for true tweet sentiment. The polarity of the seed hashtag is used to

calculate PMI [3] based association scores (Church and Hanks, 1990), substituting seed hashtags for emoticons in the technique championed by (Go et al., 2009). The lexicon contains 54,129 unigrams, 316,531 bigrams and 480,010 skip bigrams extracted from their tweet collection.

Gezi (Özdemir and Bergler, 2015) further develops this technique: nearly 20 million tweets are processed to calculate PMI scores for 376,863 unigrams, 922,773 bigrams and 850,074 dependency triples. Seed hashtags stem from 35 positive and 34 negative synonyms of *good* and *bad* in the Oxford American Writers Thesaurus (Moody and Lindberg, 2012).

We remove duplicates, retweets, and modified tweets; tweets with mixed negative and positive seed hashtags; tweets who consist mostly of URLs, hashtags, and usernames. Then, we label the tweets with the unique sentiment of their seed hashtag(s) after deleting URLs. Finally, we preprocess the collection and extract features. The tweet collection is tokenized using the CMU and Annie tokenizers (Gimpel et al., 2011; Cunningham et al., 2002), and parsed using the Stanford parser (Socher et al., 2013; de Marneffe and Manning, 2008). Negation and modality triggers are identified and their scope is determined (Rosenberg et al., 2012) in order to extract the context-aware sentiment association values[4] with PMI for unigrams, bigrams, and dependency triples (type-governor-dependent).

6 Validating Gezi on Subtask 10E

SemEval 2015 included a pilot task, Subtask 10E, which asked to determine association scores of given target terms with sentiment in tweets.

We tested our Gezi unigrams and bigrams together with the smallest but very effective aFinn lexicon in a simple rule-based approach:

1. If a target term is covered by a Gezi bigram, only this bigram score is used, to avoid double counting the unigram sentiment carrier and negation annotation, if they exist.

2. If a carrier is in a negation scope, its prior sentiment score is multiplied with -0.5.

3. Sentiment scores from aFinn and Gezi are normalized to a common scale and averaged.

4. Each prior sentiment score is scaled to [0,1]

5. If the target term cannot be assigned a score with the preceding rules, the score assigned is 0.5 (neither positive nor negative).

This approach ranked 4th among 10 submitted systems in both Kendall and Spearman correlation coefficient (Nelson, 2001) evaluations: Our Kendall rank correlation coefficient is 0.584, where other results range between 0.625 and 0.254, and our Spearman rank correlation coefficient is 0.777, where others range between 0.817 and 0.373, validating Gezi unigrams and bigrams.

7 Association Ratios to Prior Polarities

Association ratios yield continuous values and require thresholds to assign discrete prior polarities to lexicon entries.

For Gezi, we partitioned the association ratios into five categories, *strong positive, positive, exclusion, negative,* and *strong negative*. The middle category (association score close to 0) denotes terms that occur nearly as often in tweets labelled negative as in positive ones and a clear classification is not possible. The reason may be that the term is sentiment neutral (*box*) or that it can take on different sentiment in different contexts (*positive* carries negative sentiment in infection-related contexts). Rather than calling these terms *neutral*, we eliminate them entirely from Gezi.

For a term to fall into the positive categories, it has to occur at least twice as often in positive tweets as in negative tweets, thus positive terms have association scores greater than 1. For a term to be categorized as strongly positive, its score has to be greater than the geometric mean of the positive space $gMean(1, 8) = \sqrt{8} = 2.83$. Analogously, a term is considered negative if its association score lies below -1 and as strongly negative if its association score lies below -2.83. We partition the NRC Hashtag Sentiment Lexicon accordingly.

Table 3 shows the resulting composition of Gezi and the NRC lexicon for each polarity class. We see that after removing the elimination category of association scores close to 0, Gezi is roughly ten times bigger than the NRC lexicon and that the size ratio is almost equal in all the categories.

[3] pointwise mutual information

[4] Note that words have separate entries for different part-of-speech.

polar class	NRC unigrams	Gezi unigrams
strong-positive	3,390	24,739
positive	10,276	108,685
negative	8,447	62,333
strong-negative	3,605	24,639
no neutral	25,721	220,339
all	54,126	376,863

Table 3: Prior polarity class distribution.

8 Term Overlap for Different Lexica

To assess the relationship of size to unique content, we paired the five corpora and determined the size of the *Intersection* of the terms covered, indicating separately in how many cases the assigned sentiment value is the same (*Agreement*). Here, $Ratio = \frac{Agreement}{Intersection}$.

Lex A	Lex B	Intersection	Agreement	Ratio
aFinn	NRC	989	822	0.831
aFinn	Gezi	1,911	1,624	0.85
Liu	NRC	1,840	1,488	0.809
Liu	Gezi	4,028	3,386	0.841
MPQA	NRC	1,819	1,340	0.737
MPQA	Gezi	4,105	2,993	0.729
NRC	Gezi	16,868	13,957	0.827
MPQA	Liu	5,414	5,369	0.992
aFinn	Liu	1,314	1,298	0.988
MPQA	aFinn	1,246	1,202	0.965

Table 4: Intersection and agreement of lexica.

Unsurprisingly, the greatest agreement is between the smaller, manually curated lexica, which are based in part on common material (like the Harvard General Inquirer and the MPQA corpus). As expected, MPQALiu displays the greatest degree of agreement among these lexica, since MPQA formed the seed for Liu. But the lowest agreement is between MPQA and Gezi, the biggest lexicon (explained in part by the lack of the neutral category in Gezi), while Gezi-Liu has fifth-highest agreement. These observations suggest that bigger is not proportionally better: while the smaller lexica encode more of a consensus set of clear sentiment carriers, the larger lexica encode increasing amounts of low-frequency terms from the sentiment fringe, which makes their out of domain performance more volatile.

9 Experimental Setup

For the SemEval 2015 challenges, we process tweets in GATE (Cunningham et al., 2013) to extract features and run supervised machine learning algorithms using Weka (Witten and Frank, 2011).

Tokenization and POS Tagging The GATE plugin Annie tokenizer (Cunningham et al., 2002) is mature and robustly trained outside Twitter. It deals well with complex tokens, but it is not adapted to tweet-specific tokens. The CMU tokenizer (Gimpel et al., 2011) is a new tool that has been trained on Twitter data and expressly targets non-standard tokens such as emoticons, urls, exclamations (*!!!!!*), hashtags, etc. We prioritize the CMU tagger and use its tokens and POS tags when they are Twitter-specific, otherwise we use the Annie tokens, unless Ritter et al. (2011) suggests fusing multi-word entity names.

Text Normalization excludes Twitter-specific tokens that occur at the beginning and end of a sentence to improves parser performance.

Sentiment Lookup All lexical resources were transformed into gazetteer lists for each sentiment category. We use POS tag information to disambiguate senses where necessary and exclude sentiment carriers from the body of named entities.

Parsing by the Stanford parser and dependency module Version 3.4.1 (Socher et al., 2013; de Marneffe and Manning, 2008) forms the basis for NEGATOR (Rosenberg, 2013) to identify negation and modality triggers and their scope.

Feature Creation To represent our features compactly, we use compound *primary features* that encode *polarity class in linguistic context* as described above paired with the *lexical resource* that supplied this score. Abstracting away from actual sentiment terms to their polarity class helps to manage the feature space dimensionality. It also smoothes over the different lexical gaps of each lexicon. Primary features from a lexical resource are bundled under the name of that lexicon.

Table 5 shows the primary features created from the aFinn for Example 1. The only sentiment carrier term from *aFinn* is *perfect*, with a *positive* score of *3*. There is also a negation trigger which scopes over *perfect*, the scope is underlined. The resulting feature is *positive-aFinn-negated* with a score of -0.5*3=-1.5 in Table 5.

(1) El Classico on a Sunday Night is*n't* *perfect* for the Monday Morning !!

Secondary features are a collection of ad hoc features, such as specific annotations (i.e. emoticons, implicit-explicit negation triggers, modality triggers, named-entities, contrastive discourse

feature	value
positive-aFinn	0
positive-aFinn-negated	1
positive-aFinn-mod	0
positive-aFinn-mod-negated	0
negative-aFinn	0
negative-aFinn-negated	0
negative-aFinn-mod	0
negative-aFinn-mod-negated	0
aFinn-score	-1.5

Table 5: aFinn subset features for Example 1.

connectors and markers), frequencies and sentiment association scores for tokens with specific POS tags, POS tags and sentiment association scores of the first and last two tokens of tweets, and the highest and lowest sentiment association scores within tweets, see Table 6.

ids		# feat's
	Primary Feature Subsets	
f_1	aFinn	9
f_2	MPQA	12
f_3	BingLiu	8
f_4	NRC unigrams	17
f_5	NRC bigrams	17
f_6	Gezi unigrams	17
f_7	Gezi bigrams	17
f_8	dependency scores	13
f_9	dependency counts	8
	Secondary Feature Subsets	
f_{10}	POS tag based scores and counts	9
f_{11}	frequencies of specific annotations	12
f_{12}	position and top-lowest scores	6

Table 6: Feature subset bundles with IDs.

Feature Combinations We combined the twelve feature bundles of Table 6 in all possible combinations for a comprehensive ablation study. Feature combinations are processed with libSVM (Chang and Lin, 2011) with RBF kernel and parameters of cost=5, gamma=0.001 and weights=[neutral=1; positive=2; negative=2.9] in Weka (Witten and Frank, 2011) for Subtask 10B and M5P (Wang and Witten, 1997), a decision tree regressor, to predict continuous values for Task 11. These were exhaustively combined with the technique of (Shareghi and Bergler, 2013).

10 SemEval Results and Ablation

For Task 10B, tweet polarity classification, we submitted the results of $f_{1,2,3,6,8,9,10,11,12}$ containing feature subsets of aFinn, MPQA, Liu, Gezi unigrams, dependencies and secondary feature set, 94 features in total. Our submission achieved an average F-measure of positive and negative classes of 62.00, 9th among 40 submissions. Table 7 details our performance on all datasets scored. The top-performing submission achieved 64.84 f-measure. The fact that the results were this close makes it difficult to attribute them to the techniques reported wholesale and more comparison experiments need to be conducted.

dataset	F1	Rank
Twitter2015	62.00	9/40
Twitter2015Sarcasm	58.55	9/40
LiveJournal2014	73.59	6/40
SMS2013	63.05	18/40
Twitter2013	70.42	7/40
Twitter2014	70.16	9/40
Twitter2014Sarcasm	51.43	10/40

Table 7: Official results for SemEval Task 10B.

For Task 11, sentiment degree association to tweets of figurative language, we submitted $f_{1,2,3,6,7,10,11,12}$ containing aFinn, MPQA, Liu, Gezi unigrams-bigrams, and secondary feature set, totally 90 features. The challenge uses two evaluation metrics: cosine similarity and mean-squared error. According to both metrics, our submission ranked first, see Table 8.

MSE				
Overall	Sarcasm	Irony	Metaphor	Other
2.117	1.023	0.779	3.155	3.411
Cosine				
Overall	Sarcasm	Irony	Metaphor	Other
0.758	0.892	0.904	0.655	0.584

Table 8: Official results for SemEval Task 11.

Ablation studies Table 9 compares results for different feature bundles from our ablation studies. The results in italics represent official challenge results, while results in bold represent the best performing bundle for given datasets.

Our choice of system for Task 10B was informed by good performance on both, 2013 and 2014 datasets. Our best performing feature bundle is only marginally better and leaves a 2% gap to the competition winner.

For Task 11 we chose the best performing combination in 10-fold-cross validation. Our competition submission did not include dependency features. If we include them instead of MPQA and Liu feature subsets, performance increases by a cosine difference of .01.

feature ids	Task 10B F1 measures 2015	2014	2013	Task 11 Cosine
$f_{1,3,5,6,7,8,9,10,11,12}$	**62.64**	69.57	70.61	0.763
$f_{1,2,3,6,7,8,9,10,11,12}$	62.38	69.9	**70.85**	0.767
$f_{1,2,3,4,5,6,7,8,9,10,11,12}$	62.18	69.98	70.81	0.765
$f_{1,2,3,6,8,9,10,11,12}$	*62.0*	**70.16**	70.42	0.761
$f_{1,3,6,7,8,9,10,11,12}$	61.88	68.97	70.03	0.765
$f_{1,6,7,8,9,10,11,12}$	61.31	68.63	70.06	**0.768**
$f_{1,3,4,5,7,8,9,10,11,12}$	61.25	67.96	70.36	0.757
$f_{3,6,7,8,9,10,11,12}$	60.77	67.8	68.37	0.763
$f_{1,2,3,6,7,10,11,12}$	60.17	65.8	66.91	*0.758*
$f_{1,6}$	58.28	65.48	65.4	0.576
$f_{1,4}$	57.33	63.01	64.15	0.617

Table 9: Performance of different feature bundles.

No single feature bundle performs best on all datasets, however, since the best performers for each dataset include almost all features with small variations, we conclude that the different features are compatible and at least to a small degree encode complementary information. However, the feature bundle that contains all features is never the top performer, indicating some interference between features. Note that among the four top performing bundles in Table 9, only NRC unigrams is not present at all! This surprising result is probably due to Gezi being very similar but bigger, which is supported by the comparison bundles that include only aFinn and NRC or aFinn and Gezi: Gezi outperforms NRC for Task 10B by a very small margin, considering its ten-fold size difference. For Task 11, however, NRC outperforms Gezi in this baseline combination.

Table 9 shows the secondary feature bundles $f_{10,11,12}$ in every combination. These are corrective measures that were frequent and obvious enough to catch our eye and are thus very effective. More surprising is the strong performance of simple dependency feature association scores, present in all top performing feature bundles.

Impact of Size Expectedly, performing worst are the single feature bundles, in particular each lexicon used as the sole feature for the classification task, see Table 10.The surprise: aFinn, the smallest (ca. 1% of Gezi), manually curated lexicon not only dominates the others, but enhanced with our linguistic context annotations performs only 12% worse than the best bundle on 2015 data. We speculate that the reason is the design criterion (Nielsen, 2011) for aFinn to eliminate entries that may have conflicting sentiment labels altogether. This sends a very simple and clear message: reliability ranks above quantity. This of course limits aFinn to the uncontroversial core of the fuzzy set of sentiment carriers, but below that glass ceiling, it is the one to beat. Gezi, with its 100-fold size advantage trails aFinn by a mere 0.3%, which gives hope that automatically extracted lexica that include the volatile fringe can, with enough training data, approximate aFinns performance (and likely surpass it in time, as it already does for the 2014 data). The NRC lexicon which is 10-fold aFinns size, trails its performance by 5%.

feature ids	Task 10B F1 measures 2015	2014	2013	Task 11 Cosine
f_1:aFinn	54.97	60.26	62.19	0.558
f_6:Gezi uni	54.65	60.81	57.86	0.554
f_3:Liu	53.88	53.9	57.2	0.555
f_2:MPQA	52.22	51.42	53.39	0.548
f_4:NRC uni	49.83	52.39	50.9	0.609

Table 10: Task 10B's lexical sets results.

Comparing Gezi to NRC, we see that adding negation scope while enlarging the size of the tweet collection for automatically creating a resource increases its efficiency, supported by the fact that Gezi intersects and agrees with manually created lexica to a higher degree than NRC, see Table 4. But NRC outperforms Gezi in the two-lexicon-only runs $f_{1,6}$ and $f_{1,4}$ of Table 9.

11 Conclusion

Gezi, a new, large Twitter-derived sentiment lexicon that encodes the linguistic context in which a sentiment carrier occurs, was run together with certain ad hoc features on recent SemEval tasks. For comparison purposes and to improve performance, four sentiment lexica from the literature were added. A comprehensive ablation study of all the subgroupings of the resulting features shows several surprises: the smallest lexicon, aFinn, is the best solo performer. Our automatically derived Gezi lexicon marginally improves on the smaller NRC lexicon of similar design and approaches aFinns performance. We demonstrated that features do not add improvements linearly but are largely compatible with each other and effective in different subsets on different datasets, thus a careful vetting of features for each corpus is essential. Our performance in different SemEval 2015 challenge tasks shows that our approach is robust.Ranking first on the figurative language pilot task without specially geared feature additions underscores this fact and makes it a strong contender for applications across domains and tasks.

Acknowledgments

This work has been funded by a grant from Canada's Natural Science and Engineering Research Council (NSERC).

References

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1).

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. GATE: an architecture for development of robust HLT applications. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics, July 2002 (ACL'02)*, pages 168–175, Philadelphia, Pennsylvania, USA.

Hamish Cunningham, Valentin Tablan, Angus Roberts, and Kalina Bontcheva. 2013. Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. *PLoS Computational Biology*, 9(2).

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In *COLING 2008: Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation, 18-22 August 2008*, CrossParser '08, pages 1–8, Manchester, United Kingdom.

Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06, 24-26 May 2006, Genoa, Italy*, pages 417–422.

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The CoNLL-2010 shared task: Learning to detect hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning — Shared Task, July 15-16, 2010*, CoNLL '10: Shared Task, pages 1–12, Uppsala, Sweden.

Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden, and Antonio Reyes. 2015. SemEval-2015 Task 11: Sentiment analysis of figurative language in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), 1-5 June 2015*, pages 470–478, Denver, CO, USA.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2, 19-24 June 2011*, HLT '11, pages 42–47, Portland, Oregon, USA.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12.

Tobias Günther and Lenz Furrer. 2013. GU-MLT-LT: Sentiment analysis of short messages using linguistic features and stochastic gradient descent. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), June 2013*, pages 328–332, Atlanta, Georgia, USA.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 22-25, 2004*, KDD '04, pages 168–177, Seattle, WA, USA.

Alistair Kennedy and Diana Inkpen. 2005. Sentiment classification of movie and product reviews using contextual valence shifters. In *Workshop on the Analysis of Informal and Formal Information Exchange during Negotiations, FINEXIN 2005, May 26-27 2005*, Ottawa, Canada.

Alistair Kennedy and Diana Inkpen. 2006. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22(2):110–125.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 shared task on event extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task, June 4-5 2009*, BioNLP '09, pages 1–9, Boulder, CO, USA.

Svetlana Kiritchenko, Xiaodan Zhu, and Saif M. Mohammad. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research (JAIR)*, 50:723–762.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), June 14-15, 2013*, pages 321–327, Atlanta, Georgia, USA.

Rick Moody and Christine A Lindberg. 2012. *Oxford American Writer's Thesaurus*. Oxford University Press.

Roser Morante and Eduardo Blanco. 2012. *SEM 2012 shared task: Resolving the scope and focus of negation. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, 7-8 June 2012*, SemEval '12, pages 265–274, Montréal, Canada.

Roser Morante and Walter Daelemans. 2012. Annotating modality and negation for a machine reading evaluation. In Pamela Forner, Jussi Karlgren, and Christa Womser-Hacker, editors, *CLEF (Online Working Notes/Labs/Workshop), 17-20 September 2012, Rome Italy*.

Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. SemEval-2013 Task 2: Sentiment analysis in Twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), June 14-15, 2013*, pages 312–320, Atlanta, Georgia, USA.

Roger B. Nelson. 2001. Kendall Tau metric. *Encyclopaedia of Mathematics*, 3.

Finn Årup Nielsen. 2011. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages, Heraklion, Crete, Greece, May 30, 2011*, pages 93–98.

Canberk Özdemir and Sabine Bergler. 2015. CLaC-SentiPipe: SemEval2015 Subtasks 10 B, E, and Task 11. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), 1-5 June 2015*, pages 479–485, Denver, CO, USA.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10, July 6-7, 2002*, EMNLP '02, pages 79–86.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, July 27-29, 2011*, EMNLP '11, pages 1524–1534, Edinburgh, United Kingdom.

Sabine Rosenberg, Halil Kilicoglu, and Sabine Bergler. 2012. CLaC Labs: Processing modality and negation. working notes for QA4MRE pilot task at CLEF 2012. In *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012*.

Sabine Rosenberg. 2013. Negation triggers and their scope. Master's thesis, Department of Computer Science and Software Engineering, Concordia University.

Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. 2014. SemEval-2014 Task 9: Sentiment analysis in Twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), August 23-24, 2014*, pages 73–80, Dublin, Ireland.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif M. Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. SemEval-2015 Task 10: Sentiment analysis in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), 1-5 June 2015*, pages 451–463, Denver, CO, USA.

Ehsan Shareghi and Sabine Bergler. 2013. Feature combination for sentence similarity. In Osmar Zaefane and Sandra Zilles, editors, *Advances in Artificial Intelligence: 26th Canadian Conference on Artificial Intelligence, Canadian AI 2013, Regina, Canada, May 28-31, 2013*, volume 7884 of *Lecture Notes in Computer Science*, pages 150–161. Springer Berlin Heidelberg.

Richard Socher, John Bauer, Christopher D. Manning, and Andrew Y. Ng. 2013. Parsing with compositional vector grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 455–465.

Philip J. Stone, Dexter C. Dunphy, and Marshall S. Smith. 1966. *The General Inquirer: a Computer Approach to Content Analysis*. M.I.T. studies in comparative politics. M.I.T. Press, Cambridge, MA, USA.

Yong Wang and Ian H. Witten. 1997. Induction of model trees for predicting continuous classes. In *Poster in Proceedings of the 9th European Conference on Machine Learning, April 23-25 1997*, Prague, Czech Republic. Faculty of Informatics and Statistics.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, October 6-8, 2005*, HLT '05, pages 347–354, Vancouver, British Columbia, Canada.

Ian H. Witten and Eibe Frank. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers, 3rd edition.