Baekkwan Park[1] / Michael Colaresi[1] / Kevin Greene[1]

Beyond a Bag of Words: Using PULSAR to Extract Judgments on Specific Human Rights at Scale

[1] University of Pittsburgh, Department of Political Science, Pittsburgh, United States of America, E-mail: mcolaresi@pitt.edu

Abstract:
Sentiment, judgments and expressed positions are crucial concepts across international relations and the social sciences more generally. Yet, contemporary quantitative research has conventionally avoided the most direct and nuanced source of this information: political and social texts. In contrast, qualitative research has long relied on the patterns in texts to understand detailed trends in public opinion, social issues, the terms of international alliances, and the positions of politicians. Yet, qualitative human reading does not scale to the accelerating mass of digital information available currently. Researchers are in need of automated tools that can extract meaningful opinions and judgments from texts. Thus, there is an emerging opportunity to marry the model-based, inferential focus of quantitative methodology, as exemplified by ideal point models, with high resolution, qualitative interpretations of language and positions. We suggest that using alternatives to simple bag of words (BOW) representations and re-focusing on aspect-sentiment representations of text will aid researchers in systematically extracting people's judgments and what is being judged at scale. The experimental results below show that our approach which automates the extraction of aspect and sentiment MWE pairs, outperforms BOW in classification tasks, while providing more interpretable parameters. By connecting expressed sentiment and the aspects being judged, PULSAR (Parsing Unstructured Language into Sentiment-Aspect Representations) also has deep implications for understanding the underlying dimensionality of issue positions and ideal points estimated with text. Our approach to parsing text into aspects-sentiment expressions recovers both expressive phrases (akin to categorical votes), as well as the aspects that are being judged (akin to bills). Thus, PULSAR or future systems like it, open up new avenues for the systematic analysis of high-dimensional opinions and judgments at scale within existing ideal point models.

Keywords: text analysis, Human Rights, Natural Language Processing

DOI: 10.1515/peps-2018-0030

1 Introduction

While there has been important quantitative work on measuring sentiment, judgments and positions in international relations and across the social sciences, this research has traditionally used statistical models of discrete, lower-dimensional data such as categorical responses from public opinion polls and recorded votes, not political text. Conversely, qualitative research has long relied on the patterns in texts to understand nuanced trends in public opinion, the terms of international alliances, the demands of dissatisfied groups, and the positions of politicians (Ho & Quinn, 2008; Monroe & Maeda, 2004). Yet, qualitative human reading does not scale to the accelerating mass of digital information available currently. Researchers are in need of automated tools that can extract meaningful opinions and judgments from texts. Thus, we need to marry the model-based, inferential focus of quantitative social science, as exemplified by ideal point models, with high resolution, qualitative interpretations of language and positions.

The richness of human language has meant that most researchers do not think of text as data at all, but only as a conduit for measures and summaries that must be extracted by humans. Thus, the quantitative use of unstructured text to understand sentiment and opinions has traditionally been less prevalent than other less complex data, such as dichotomous votes or membership choices. Exceptions include event coding (Brandt, Freeman, & Schrodt, 2014; Schrodt, Beieler, & Idris, 2014) and topic modeling (Blei, Ng, & Jordan, 2003; Quinn, Monroe, Colaresi, Crespin, & Radev, 2010), neither of which attempts to measure opinions, sentiments or judgments of individuals directly. Other exceptions that do attempt to measure judgment or positions include one-dimensional sentiment analysis (Liu, 2015) as well as text scaling applications (Laver, Benoit, & Garry, 2003; Lowe, 2013; Monroe & Maeda, 2004; Slapin & Proksch, 2008). These latter research areas provide clues for how to move current practices forward.

Michael Colaresi is the corresponding author.

We suggest that using alternatives to the conventional bag of words (BOW) representations and re-focusing on aspect-sentiment representations of text will aid researchers in systematically extracting people's judgments and what is being judged at scale. The experimental results below show that our approach which automates the extraction of aspect and sentiment phrase pairs, outperforms BOW in classification tasks, while providing more interpretable parameters.

1.1 The bag of words (BOW) representation of text and scrambled sentiment

Representing a textual document as a bag-of-words implies that the terms within that document are simply counted and entered into a vocabulary-length vector. Although the BOW assumption can be justified on computational efficiency grounds and has been used widely, it has also been criticized due to its unrealistic assumption about the (lack of) dependencies between words (Wallach, 2006). As Wallach (2006) writes, examples such as "the department chair couches offers" and "the chair department offers couches" are identical in terms of unigram BOW statistics, but they have very different semantic meanings when read. There are strong a priori reasons to expect that word order, the grammatical role a word plays, and syntax can matter for communicating sentiment and judgments.

The problems with using BOW representations to extract sentiment are more obvious when applied to economic and political examples. The statement "You should buy the iPhone, not a Samsung phone" has the same BOW representation as "You should buy a Samsung phone, not the iPhone", but conveys different sentiment towards the iPhone. Similarly, this example highlights the fact that the placement of negation matters a great deal for the expressed sentiment. In the human rights context, the two statements, "Laws protect public accountability from official corruption", and "Laws protect official corruption from public accountability" express very different valences concerning the law.

A growing literature on supervised and rule-based learning approaches to sentiment analysis have also reached the conclusion that it is propitious to move beyond BOW representations. Building on the work of Pang, Lee, and Vaithyanathan (2002) researchers have begun adding non-BOW features such as negation, intensifiers, grammatic parsing and syntactic dependences, presenting systems that substantially improve upon the baseline accuracy and usefulness of BOW-based models (Feldman, 2013; Liu, 2015; Pang & Lee, 2008).[1] Some of the most impressive recent results have come from research that attempts to extract not only the aggregate judgment from a sentence or paragraph, but also identify explicitly what is being judged (see Liu (2015)).

2 Our approach: from ABSA to PULSAR

Aspect-based Sentiment Analysis (ABSA) builds on conventional sentiment analysis to explicitly identify the aspects of given target entities and estimates the judgment for each mentioned aspect (Liu, 2015). While this research has made an impact in the natural language processing domain, it has not yet been used in the social sciences or IR research to our knowledge. One potential reason ABSA-tasks have been ignored to date is that it necessitates non-BOW representations of text. The widespread use of BOW representations, either implicitly (e.g. dictionary-based approaches) or explicitly (e.g. some supervised learning and scaling applications), disconnects expressed sentiment from the aspect being judged. From the perspective of ABSA, BOW representations assume that words express the same sentiments and opinions regardless of the other words that appear with them in a document. This assumption is violated not only in cases of negation and adjectival/adverbial modification, but wherever speakers are discussing distinct issues and aspects. The political judgment word "support" maps to very different latent positions depending on whether it modifies aspects of international politics such as "intervention"/"state-sovereignty or "civil liberties"/"torture".

2.1 Parsing unstructured language into sentiment-aspect representations (PULSAR)

We suggest that ABSA presents a framework for two important roles that phrases can play within a sentence, sentiment-expression and aspect-expression. Here we introduce and explore the performance of a tool we have built that can extract/tag phrases that communicate these distinct modalities, as well as link them together. We deviate from previous ABSA research in two respects. First, where ABSA research has concentrated on product and movie reviews, our focus is on building an open-source tool that will be useful to understand human rights and other social science topics through improved models, prediction and scaling. Second, we are interested not only in the overall performance of systems that use sentiment-aspect expressions, but their interpretability. One of the benefits of using text as data is that the parameters can be compared to qualitative semantic judgments based on reading a subset of the texts. It is important for research tools to facilitate the process by which researchers learn the strengths and weaknesses of their model (Colaresi & Mahmood, 2017).

As an intermediate step towards these ends, we introduce PULSAR, a tool that processes natural language corpora into structured aspect-based sentiment expressions. The three main tasks of PULSAR are 1) extracting aspects, 2) extracting sentiment, and 3) linking the corresponding sentiment and aspect expressions. PULSAR 1.0 takes a syntactic approach to these three tasks. We have designed rules about grammar dependency relations between opinion-oriented words and aspects to guide the identification of phrases, their modalities, and links between them (Liu, Gao, Liu, & Zhang 2013; 2015; Qiu, Liu, Bu, & Chen, 2011; Wu, Zhang, Huang, & Wu, 2009).As such, PULSAR relies on word order, automated grammar tagging, and automatic syntax parsing, which we detail below.

PULSAR includes a sub-routine that takes domain-specific multi-word expressions (MWE) as input, but can also return frequent MWEs and what role they are used in (aspect or sentiment). A MWE is a sequence of words that acts as a single unit at some level of linguistic analysis (Calzolari et al., 2002). “Human rights” is itself a MWE as is “the rule of law”. MWEs have played an important role in the description and processing of natural language processing tasks, but have been underutilized in IR and the social sciences. The BOW approach does not preserve meaningful multi-word phrases (Handler, Denny, Wallach, & O’Connor, 2016). Even when using bigrams and trigrams, which create a large number of meaningless parameters from contiguous unigrams, MWEs that are longer than the the n-gram’s preset limit will be missed.

2.2 Extracting aspect and sentiment: matching strategies

Identifying meaningful sequences of words (the aspect expressions and sentiment information) is a difficult set of tasks (Sag, Baldwin, Bond, Copestake, & Flickinger, 2002). The current version of PULSAR, builds on previous work from the NLP community, particularly the open-source Stanford CoreNLP software (Manning et al., 2014), which allows us to define sets of syntactic patterns based on parts of speech tags, and syntactic dependencies in the text, along with their dictionaries, to identify the aspect-based sentiment expressions.

Aspects Aspects are largely noun phrases, thus we begin by combining part-of-speech (POS) tagging with a set of rules for defining aspect MWEs. We rely heavily on Justeson and Katz (1995)’s MWE extraction method which uses a set of POS patterns to find and extract noun phrases. Our pattern-based method is specified in terms of two parameters (G_A, M_A). Our aspect grammar G_A is a non-recursive regular expression that defines a set of POS tag sequences. Importantly, and in contrast to past work using n-grams, we do not specify the length of a phrase for any sentence. For example, a noun preceded by several adjectives would fit our rule (described in more detail below). Our aspect matching strategy M_A, defines how we scan a document to apply G_A. For this process, we begin scanning from the end of a given sentence and identify noun phrases including one or more adverbs (POS tags R: RB, RBR, RBS), adjectives (POS tags A: JJ, JJR, JJS), nouns (POS tags N: NN, NNS), excluding prepositions or proper nouns. We can express this as a Perl-like regular expression:

$$(\mathrm{R}|\mathrm{A})|((\mathrm{A}|\mathrm{N}) * \mathrm{N})$$

For a concrete example, take the sentence, “*There were no politically motivated disappearances.*”. PULSAR would first use a POS tagger to identify the grammatical role each token plays in the sentence. The output of this step is provided in Figure 1.

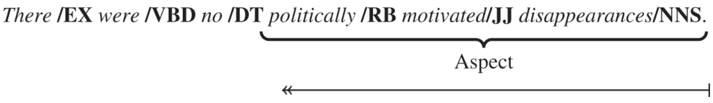
There /EX were /VBD no /DT politically /RB motivated/JJ disappearances/NNS.
Aspect

Figure 1: POS tagging within PULSAR to extract aspects. This examples uses the Stanford CoreNLP toolchain (Manning et al., 2014).

From this POS tagging (Toutanova, Klein, Manning, & Singer, 2003), our aspect-extraction regular expression only matches the phrase [*politically_motivated_disappearances*], since disappearances is preceded by one adjective and then one adverb. The noun (NNS) is the root of this phrase it is classified as an aspect MWE of the sentence.

Sentiment In order to identify sentiment MWEs, our pattern-based method for sentiment is specified in terms of two parameters (G_S, M_S). G_S is our sentiment grammar that defines a set of the *Penn Treebank* syntactic tagsets (Taylor, Marcus, & Santorini, 2003). Our sentiment matching strategy M_S defines how we scan a sentence to apply G_S. For the sentiment expression, we take advantage of the *Penn Treebank* syntactic tree structure and develop a second grammar to capture them. After parsing a sentence for its syntactic structure presented in the *CoNLL-X* format of the tree bank, we start searching syntactic tagsets from the beginning of the sentence and capture all verb phrase syntactic tagsets (Syntactic tag VP) with negation *no* (POS tag ND: DT), if any. Because negation changes the direction (meaning) of sentiment, we emphasize it by adding a *NEG* tag to the sentiment expression. If negation is absent we define the verb-phrases through a set of patterns that are similar to the aspect regular expressions.

$$(\mathrm{ND}|\mathrm{VP})*$$

If we take the same sentence above, Figure 2 shows the *Penn Treebank Tree* structure of the sentence.

```
                (ROOT
                  (S
                    (NP    (EX   There))
             ⎧       (VP     (VBD    were )
Sentiment    ⎨      (NP      (DT       no )
             ⎩
                      (ADJP     (RB politically))   (JJ motivated))
                      (NNS     disappearances )))
                    (.   .)))
```

Figure 2: Penn tree for extracting sentiment. This examples uses the Stanford CoreNLP toolchain (Manning et al., 2014).

From this we capture [*NEG_were*] as the MWE based sentiment of the sentence.

Pairing the Aspect with the Sentiment Once we have captured the expressions of aspects and sentiments, the next task is to identify and pair the sentiment expression with the extracted aspects. The left illustration in Figure 3 is the original dependency parsing results. It does not recognize the extracted MWE aspect above. The right illustration in Figure 3 shows the dependency relations between the extracted aspect and the corresponding sentiment in the sentence. PULSAR searches for syntactic dependencies between an aspect MWE and sentiment MWE in the same sentence. If a specific dependency is found, a pair is formed. In our example, PULSAR outputs the aspect-sentiment pair [(*politically_motivated_disappearances, NEG_were*)], where NEG is our tag for negation.

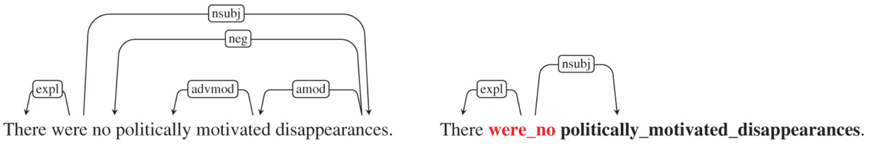
nsubj
neg
expl
advmod
amod
There were no politically motivated disappearances.
expl
nsubj
There were_no politically_motivated_disappearances.

Figure 3: Dependency between Aspect and Sentiment.

Thus, PULSAR uses POS tagging which depends on word order, and then POS tagging to measure syntactic dependences between words within identified MWEs. The sentiment-aspect pairs can then be used as new tokens and counted and quantified for supervised or unsupervised learning tasks. In this and other examples, PULSAR is able to differentiate negation, distinct sentiment on the same expressed aspect, as well as identify when the same sentiment is offered on disparate aspects.

3 Experiments: using US state department human rights reports to predict the political terror scale

We apply PULSAR to *the annual US State Department Country Reports on Human Rights Practices* in order to highlight how aspect-based sentiment features represent human rights over time,[2] as well as explore whether we can more accurately forecast the human coded Political Terror Scale (PTS) (Gibney, Cornett, Wood, Haschke, & Arnon, 2015) with aspect-based sentiment features as compared to BOW-generated features. Additionally, we want to compare the interpretability of the parameters between the two approaches.[3]

3.1 Results

Our results, comparing the out-of-sample accuracy of predictions of a state's human rights score (PTS) using bag-of-word features and PULSAR's sentiment-aspect features, are presented in Table 1. We train each model on the text and PTS scores from 1999–2008, and test the accuracy of each model's predictions for the PTS scores in 2009–2010. We use four different model representations in these sets of experiments, naive Bayes (NB), logistic regression (LR), a support vector machine with a radial basis function (SVM), and a random forest (RF), each first using only BOW features and next using PULSAR-generated features. Note that the labels have 5 classes so the baseline random accuracy is .20.

Table 1: Out of sample accuracy of bag-of-words (BOW) features and sentiment-aspect pairs (S-A pairs) across four models, Naive Bayes, Logistic, SVM, and Random Forest.

BOW Features	Accuracy	S-A Features	Accuracy
BoW (NB)	0.66	SA Pairs (NB)	0.71
BoW (LR)	0.67	SA Pairs (LR)	0.72
BoW (SVM)	0.66	SA Pairs (SVM)	0.70
BoW (RF)	0.67	SA Pairs (RF)	0.70
Baseline	0.20	Baseline	0.20

Results from our experiments predicting PTS scores with BoW features (on the left) and PULSAR output (on the right). Across different supervised learning algorithms (rows), input from the PULSAR sentiment-aspect pairs (SA) improves the ability for systems to accurately predict the human coded-judgments in the PTS data.

We find that using the PULSAR-generaed sentiment-aspect pairs boosts out-of-sample accuracy, relative to BOW features, by approximately 5 percent. This is despite the fact that the labels in this case are applied to the document as a whole, and not specific sentences. Issue-specific labels, such as CIRI Physical Integrity scores in Cingranelli, Richards, and Clay (2014) may show even greater gains. This result supplies some useful evidence that when researchers are interested in extracting opinions and judgments from text, the added complexity of taking syntax and grammar into account may be worth it.

We also find that the features that shift predictions most dramatically are more interpretable across the classes for the sentiment-aspect pairs that PULSAR creates, as compared to BOW features. While the BOW features in Table 2 provide some hints to the underlying patterns, many proper nouns are included in the top words. This fact suggests that these models are overlearning names, and not generalizable features that would apply to new countries (like South Sudan). There are also some curious words, such as "province" and "south", being top features for the worst category (5). In fact, to make sense of the top BOW features, we often must place them in an implied phrase, which may or may not be accurate. For example "civil" is a top feature for category (2), a relatively good score. However, this could refer to civil war, civil rights or civil cases. The feature itself does not communicate which meaning or context is relevant here.

Table 2: Bag-of-words top features, based on feature weight, predicting labels 1 (Best) to 5 (Worst), vs. all other categories.

1 (Best)	2	3	4	5 (Worst)
provides	generally	lebanese	2004	civilians
right	civil	saudi	moscow	province
law	moldova	mexico	developments	forces
respects	detention	section	drc	south
1999	legal	militants	ordinance	paramilitary
council	documents	officials	martial	killed
act	indonesia	mob	baha	rebel
european	indonesian	election	china	iraqi
generally	code	moj	killings	guerrillas
percent	shari	ghana	members	korean

The terms that have the highest weights in predicting the best to worst PTS scores. It is difficult to decipher why these terms would systematically lead to better/worse PTS scores, especially as proper nouns are prominently included.

By comparison, the sentiment-aspect pairs in Table 3 are much clearer. The parser has cleared most specific named entities, to focus on abstract aspects, such as "human rights" and the treatment of "civilians", "women", and "children". The sentiments are also able to capture whether there were "reports" or not, with "neg" representing negation. The top features also progress from "respect,govern" (1) to "kill, civilians" and other groups (5). The features in the middle categories capture new nuances that would be impossible to communicate with only single words. If there "remain, problems" but countries "generally respect , human rights", the score is likely to be a 2. If there "remain , seriou problem", and "killing, (by) security forces" is reported, then the score is likely to be a harsher 4. Thus, we find that the PULSAR generated sentiment-aspect pairs provide improved out-of-sample performance and greater interpretability. In this sense, PULSAR provides features from the text that help us understand what patterns the models are learning.

Table 3: Aspect-sentiment pair top features, based on feature weight, predicting labels 1 (Best) to 5 (Worst), vs. all other categories.

1 (Best)	2	3	4	5 (Worst)
respect, govern	were, problem	wa, serious_problem	neg_were, develop	kill, civilian
permit, visit	gener_respect, human_right	were, previou_year	remain, seriou_problem	kill, person
employ, govern_offici	respect_gener, govern	were, instanc	wa_taken_against, action	kill, soldier
gener_enforc, right	ha, author	use, tear_ga	kill, civilian	result_in, death
gener_observ, prohibit	gener_oper_without, govern_restrict	use, govern	remain, problem	see, section
respect, right	remain, problem	were, seriou_problem	continu, govern	rape, women
provid, law	were, human_right	beat, polic	kill, secur_forc	includ, children
effect_enforc, govern	provid_for, constitut	cooper_with, govern	arrest, polic	kill, secur_forc
gener_observ, govern	wa, problem	remain_pend_at, year_end	accord_to, govern	took, action
neg_were, report	were, such_practic	restrict, right	remain_in, prison	includ, kill

The SA pairs that have the highest weights in predicting the best to worst PTS scores. It is much clear to identify why documents with these terms would likely receive these scores. There are many fewer proper nouns, and negation is featured where it would be expected. For example, ”neg_were, report” represents the phrases that indicate the lack of reports of violations.

4 Conclusion

In this paper, we introduce PULSAR, a user-friendly tool that helps social science researchers to convert large-scale natural language corpora into structured aspect-sentiment expressions. PULSAR can aid IR research in particular by removing dependencies on limited bag-of-words and bag-of-ngrams representations of text, a necessary step for more accurately extracting sentiment from speech. We show that this approach can improve prediction based tasks on human rights specifically, while providing more interpretable parameters.

By connecting expressed sentiment and the aspects being judged, PULSAR has deep implications for understanding the underlying dimensionality of issue positions and ideal points estimated with text. Monroe and Maeda (2004) make clear that BOW-scaling is unable to simultaneously estimate the positions and difficulty parameters of words, since we only have the counts of words, not their counts across different contexts. This limitation arises because word-scaling models treat every situation that leads to a word utterance, as identical, leading researchers to often subset their analysis by a given topic (Monroe, Colaresi, & Quinn, 2008). Our approach to parsing text into aspects-sentiment expressions recovers both expressive phrases (akin to categorical votes), as well as the aspects that are being judged (akin to bills). Thus, PULSAR or future systems like it, open up new avenues for the systematic analysis of high-dimensional opinions and judgments at scale within existing ideal point models.

Notes

[1] With the growing interest in neural networks (deep learning), scholars have begun to incorporate the continuous representations of words as features and shown improvement in capturing sentiment (Bespalov, Bai, Qi, & Shokoufandeh, 2011; Socher, Lin, Manning, & Ng, 2011; Socher, Huval, Manning, & Ng, 2012; 2013; Yessenalina & Cardie, 2011).

[2] https://www.state.gov/j/drl/rls/hrrpt/

[3] In our experiment stemming was used on the PULSAR features.

References

Bespalov, D., Bai, B., Qi, Y., & Shokoufandeh, A. (2011). Sentiment classification based on supervised latent n-gram analysis. *Proceedings of the 20th ACM international conference on Information and knowledge management – CIKM '11*. URL: http://dx.doi.org/10.1145/2063576.2063635

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Brandt, P. T., Freeman, J. R., & Schrodt, P. A. (2014). Evaluating forecasts of political conflict dynamics. *International Journal of Forecasting*, *30*(4), 944–962.

Calzolari, N., Fillmore, C. J., Grishman, R., Ide, N., Lenci, A., MacLeod C., & Zampolli, A. (2002). Towards best practice for multiword expressions in computational lexicons. In *LREC*.

Cingranelli, D. L., Richards, D. L., & Clay, K. C. (2014). *The CIRI human rights dataset*. v.2014.04.14.

Colaresi, M., & Mahmood, Z. (2017). Do the robot: Lessons from machine learning to improve conflict forecasting. *Journal of Peace Research*, *54*(2), 193–214.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, *56*(4), 82–89. URL: http://dx.doi.org/10.1145/2436256.2436274

Gibney, M., Cornett, L., Wood, R., Haschke, P., & Arnon, D. (2015). The political terror scale 1976–2015. Date Retrieved, from the Political Terror Scale website: http://www.politicalterrorscale.org.

Handler, A., Denny, M., Wallach, H., & O'Connor, B. (2016). Bag of what? Simple noun phrase extraction for text analysis. In *Proceedings of the First Workshop on NLP and Computational Social Science*. pp. 114–124.

Ho, D. E., & Quinn, K. M. (2008). Measuring explicit political positions of media. *Quarterly Journal of Political Science*, *3*(4), 353–377.

Justeson, J. S., & Katz, S. M. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, *1*(1), 9–27.

Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review*, *97*(1), 311–331.

Liu, B. (2015). *Sentiment Analysis: Mining Opinions, Sentiments and Emotions*. New York: Cambridge University Press.

Liu, Q., Gao, Z., Liu, B., & Zhang, Y. (2013). A logic programming approach to aspect extraction in opinion mining. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences*. Vol. 1 IEEE pp. 276–283.

Liu, Q., Gao, Z., Liu, B., & Zhang, Y. (2015). Automated rule selection for aspect extraction in opinion mining. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Lowe, W. (2013). There's (basically) only one way to do it. Available at SSRN.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*. pp. 55–60. URL: http://www.aclweb.org/anthology/P/P14/P14-5010

Monroe, B. L., & Maeda, K. (2004). Talk's cheap: Text-based ideal point estimation. In *presented to the Political Methodology Society*. Palo Alto, CA.

Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). 'Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, *16*(4), 372–403.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Informational Retrieval*, *2*(1-2), 1–135.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? *Proceedings of the ACL-02 conference on Empirical methods in natural language processing – EMNLP '02*. URL: http://dx.doi.org/10.3115/1118693.1118704

Qiu, G., Liu, B., Bu, J., Chen, C. (2011). Opinion word expansion and target extraction through double propagation. *Computational Linguistics*, *37*(1), 9–27.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M., & Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, *54*(1), 209–228.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: a pain in the neck for NLP. In *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer pp. 1–15.

Schrodt, P. A., Beieler, J., & Idris, M. (2014). Three'sa charm?: Open event data coding with el: Diablo, Petrarch, and the open event data alliance. In *ISA Annual Convention*.

Slapin, J. B., & Proksch, S.-O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, *52*(3), 705–722.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*. pp. 1631–1642.

Socher, R., Huval, B., Manning, C. D., & Ng, A. Y. (2012). Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*. Association for Computational Linguistics pp. 1201–1211.

Socher, R., Lin, C. C., Manning, C., & Ng, A. Y. (2011). Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th international conference on machine learning (ICML-11)*. pp. 129–136.

Taylor, A., Marcus, M., & Santorini, B. (2003). The Penn treebank: an overview. In A. Abeillé (Ed.), *Treebanks* (pp. 5–22). Dordrecht: Springer.

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics pp. 173–180.

Wallach, H. M. (2006). Topic modeling. *Proceedings of the 23rd international conference on Machine learning – ICML '06*. URL: http://dx.doi.org/10.1145/1143844.1143967

Wu, Y., Zhang, Q., Huang, X., & Wu, L. (2009). Phrase dependency parsing for opinion mining. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Vol. 3 Association for Computational Linguistics pp. 1533–1541.

Yessenalina, A., & Cardie, C. (2011). Compositional matrix-space models for sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP ′11 Stroudsburg, PA, USA: Association for Computational Linguistics pp. 172–182. URL: http://dl.acm.org/citation.cfm?id=2145432.2145452