Towards Quantifying the Distance between Opinions

Saket Gurukar,[1]* Deepak Ajwani,[2] Sourav Dutta,[3] Juho Lauri, Srinivasan Parthasarathy,[1] Alessandra Sala[4]
[1]Department of Computer Science and Engineering, The Ohio State University
[2]School of Computer Science, University College Dublin
[3]Huawei Ireland Research Center
[4]Nokia Bell Labs
gurukar.1@osu.edu, deepak.ajwani@ucd.ie, sourav.dutta2@huawei.com,
juho.lauri@gmail.com, srini@cse.ohio-state.edu, alessandra.sala@nokia-bell-labs.com

Abstract

Increasingly, critical decisions in public policy, governance, and business strategy rely on a deeper understanding of the needs and opinions of constituent members (e.g. citizens, shareholders). While it has become easier to collect a large number of opinions on a topic, there is a necessity for automated tools to help navigate the space of opinions. In such contexts understanding and quantifying the similarity between opinions is key. We find that measures based solely on text similarity or on overall sentiment often fail to effectively capture the distance between opinions. Thus, we propose a *new distance measure* for capturing the similarity between opinions that leverages the nuanced observation – *similar opinions express similar sentiment polarity on specific relevant entities-of-interest*. Specifically, in an unsupervised setting, our distance measure achieves significantly better Adjusted Rand Index scores (up to 56x) and Silhouette coefficients (up to 21x) compared to existing approaches. Similarly, in a supervised setting, our opinion distance measure achieves considerably better accuracy (up to 20% increase) compared to extant approaches that rely on text similarity, stance similarity, and sentiment similarity.

Introduction

Crucial decisions in public policy-making as well as in business strategy can be enhanced through a deeper understanding of the diverse opinions and perspectives put forth by relevant stakeholders. While elections, referendums, and market surveys provide important mechanisms for gauging public opinion, in general, they are (i) expensive and (ii) primarily involve selecting from a predefined set of options. Furthermore, many decision processes (i) may not justify the huge expense of referendums, and might also (ii) require a more nuanced analysis of opinions on a more frequent or regular basis. This has led to the growth of *digital democracy platforms* for continuous collection of public opinions at a significantly less cost, enabling better analysis and alignment of decisions with the viewpoints of the stakeholders. For instance, governing institutions in many democratic countries issue public notices to seek opinions on governing policies (e.g., Net neutrality NPRM issued by the U.S. Federal Communications Commission or the Brexit referendum). On a smaller scale, even local city councils call for public consultations on administrative issues such as local property taxes or road works. Similarly, businesses spend a considerable amount of resources to understand and organize customer feedback on products and services.

In many cases, the collection of opinions is too large to be manually curated. For instance, on the BBC News website, popular articles receive thousands of comments. Thus, there is a need for automated tools to not only navigate but also assist in understanding the space of all opinions. A fundamental challenge in navigating opinions (or clusters of opinions) is the need to construct a distance measure that *quantifies the distance between opinions*. A good distance measure should be able to semantically differentiate between opinions that are highly similar and opinions that are opposing. Despite the fundamental importance of such an opinion distance function, we note that the traditional approaches are inadequate for this purpose.

Many existing approaches for *opinion mining* rely on features based on the text-similarity of the opinion documents (Mullen and Collier 2004) or the overall sentiment orientation of the comment (i.e., whether the overall tone is positive, negative or neutral) (Pang, Lee, and Vaithyanathan 2002). Unfortunately, such approaches involving text-similarity and sentiment analysis are often inadequate in our problem setting. For instance, consider the diametrically opposite opinions of `"In this debate, Hillary looked presidential while Trump came across as manipulative"` and `"In this debate, Trump looked presidential while Hillary came across as manipulative"`. Use of text-similarity or overall sentiment based features will categorize the above opinions to be **very similar** as they have the same bag-of-words, similar sentence structure and identical set of sentiment words. In fact, we demonstrate later that text-similarity based features like TF-IDF and also semantic measures like Word-Mover distance and Doc2vec are poor indicators of opinion similarity on many datasets. Furthermore, such features resulted in opinion clusters that

*This work was initiated when the first three authors were at Nokia Bell Labs, Ireland

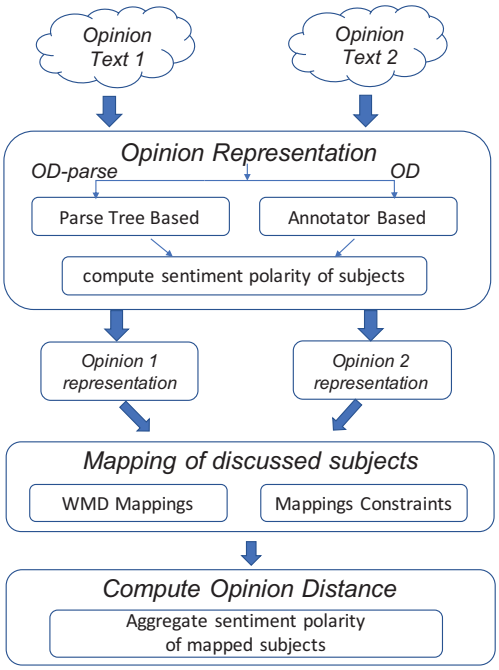

Figure 1: Opinion Distance Pipeline

have very little agreement with the ground truth clusters, even for cases with only two clusters.

The above exemplar suggests that to effectively capture the differences between opinions, a measure needs to be able to understand deeper nuances of the semantics of the text. Given the inherent difficulties in capturing the "true" opinion distance (as opposed to text-similarity distance), recent research has focused on a simpler variant of this problem called *stance detection* (i.e., whether an opinion is in-favour of, neutral or in-opposition to a target topic). However, human opinions are often nuanced and organizing them independently into stances can lead to distortion or misinterpretation – a scenario increasingly crucial in the political arena. For example, on the issue of Brexit, consider the following opinions: (i) "Brexit will result in economic loss, no doubt, but because it will reduce immigration numbers drastically, it will still be worth it"; and (ii) "Brexit will result in economic prosperity and huge savings and both the quality and quantity of the immigrants will increase". Both are in favour of Brexit (depicting the same stance), but are still fundamentally different in terms of the opinions as to why the same stance is supported. Existing similarity measures fail to capture such nuanced opinion analysis, which might result in wrong conclusions about what people really opine about a topic.

Contributions: We address the problem of opinion clustering by proposing a distance measure that quantifies the semantic similarity between opinions. Leveraging the observation that similar opinions express similar sentiment polarity on the relevant subjects, we represent an opinion in terms of the discussed subjects and the sentiment polarity expressed towards those subjects.[1] The subjects in the opinion texts are then mapped based on semantic similarity and the opinion distance is defined as an aggregate of sentiment polarity difference towards the mapped subjects. We study a few concrete instantiations of our distance measures and propose a carefully engineered computational pipeline (c.f. Figure 1). We also demonstrate improved experimental results (in both supervised and unsupervised settings) on several real-world data-sets along with a use-case study to organize comments on BBC news portal, showcasing the efficacy of the proposed distance measures for organizing opinions over extant approaches on both unsupervised and supervised task settings.

Background

We begin by briefly reviewing the **key definitions**:

1. *Target Entity*: The subject or topic of interest that frames the discussion or narrative.
2. *Aspect*: Characteristics of the target entity.
3. *Perspective*: A way of viewing or perceiving the target entity or topic by a person.
4. *Stance*: Subjective disposition towards the topic, typically, characterized as being in favor, neutral towards, or against defined target entity.
5. *Opinion*: The statement(s) reflecting and/or justifying the belief or judgment of a person towards a target entity or its aspect(s). Note that *two opinions may be very different, even though they may have the same stance towards the target entity*.

Related work

Contrastive opinion modeling (Fang et al. 2012) relies on computing topic models – a variant of LDA using Gibbs sampling to estimate the model parameter. The authors subsequently adopt the Jensen-Shannon divergence among the individual topic-opinion distributions to determine contrastive opinions.[2] The problem of identifying different perspectives or viewpoints about a topic is addressed by proposing a graph partitioning method which exploits the presence of *social interactions* in order to identify viewpoints (Quraishi, Fafalios, and Herder 2018). Our current effort does not rely on social interactions in order to quantify the opinion distance. The Author Interaction Topic Viewpoint model (AITV) (Trabelsi and Zaïane 2018) focuses on the task of viewpoint detection of author and post. The AITV topic model levers authors' interactions and encodes the heterophily assumption that difference in viewpoints induces more interactions. However, such interactions are not always available.

Stance detection has been widely studied with a focus on short-text social media (Mohammad et al. 2016) or news articles (Stab and et al. 2018; Riedel et al. 2017; Awadallah,

[1] A similar representation of opinion is considered in (Kim and Hovy 2006; Kobayashi et al. 2004; Recupero et al. 2015; Liu, Hu, and Cheng 2005).

[2] We could not obtain the source code.

Ramanath, and Weikum 2012). Many studies (Riedel et al. 2017) have shown that features based on n-grams trained with SVM are difficult-to-beat baselines for such tasks. Similarly, sentiment analysis often rely on dependency parse trees (Socher et al. 2013; Zhang, Wang, and Liu 2018) and aspect-based sentiment analysis (Pontiki and et al. 2016; Pavlopoulos 2014; Titov and McDonald 2008) using conditional random field classifiers have also been proposed. However, note that *stance*, *sentiment*, and *opinion* have nuanced differences as defined earlier.

Sentiment analysis has been used in customer feedback on a brand name. Socher et al. (Socher et al. 2013) compute the sentiment of a sentence by assigning sentiment to individual words and phrases in the dependency parse tree of the sentence and then recursively aggregate the sentiments in the parse tree to compute the sentiment of the sentence. For more on sentiment analysis, see (Buche, Chandak, and Zadgaonkar 2013; Pang, Lee, and others 2008; Liu and Zhang 2012). Aspect-Based Sentiment Analysis (ABSA) (Pontiki and et al. 2016; Pavlopoulos 2014; Titov and McDonald 2008) is a subfield of sentiment analysis where sentiments towards each aspect are studied. However, most of the studies comprises of supervised methods and require the definition of aspects to be known beforehand. Note that, in this work, we focus on developing an unsupervised distance measure for opinions.

There has been considerable work (e.g., (Kim and Hovy 2006)) on extracting opinion targets and expressions. However, most of the proposed methods are supervised and domain specific (Wiegand, Schulder, and Ruppenhofer 2015). The Sentilo tool (Recupero et al. 2015) identifies the discussed entities in opinion and the sentiment expressed towards the entity in an unsupervised and domain-independent way. However, the methods proposed in opinion target and opinion expression extraction literature cannot be extended straightforwardly to compute opinion distance. For instance, the issue of opinion subject polysemy while computing opinion distance is nontrivial to solve.

We stress that the problems addressed in the above studies are different than ours. To our knowledge, there is no work on an "opinion distance measure" focusing on quantifying the similarity or dissimilarity between different opinions.

Distance measures

For designing a distance measure for opinions, we leverage the observation that similar opinions express similar sentiment polarity on the relevant discussed subjects. We propose a set of measures based on aggregating the difference in sentiment polarity of words associated with the subjects.

The key to designing a good distance measure lies in the representation of the opinion itself. That is, instead of merely relying on the overall tone/sentiment of the opinion or just considering the bag-of-words or collection of n-grams representations, we propose a more nuanced representation of opinion in terms of the discussed subjects and the sentiment polarity expressed towards those subjects. To this end, our distance computation framework: (i) extracts the opinion subjects discussed in the text, (ii) identifies the words associated with the subjects that express opinions towards those subjects, and (iii) computes the sentiment polarity of the associated words.

Given two opinions (O_1 and O_2), each represented in terms of a tuple of opinion subjects and sentiment polarity towards the corresponding subject within each opinion, their distance is computed by: (i) first, mapping opinion subjects in O_1 to corresponding subjects in O_2 (and vice versa) based on their semantic similarities; and (ii) second, aggregating the difference of sentiment polarities expressed on the correspondingly mapped subjects across both opinions. Figure 1 presents the different steps in our pipeline. Note that it is crucial to have a reasonably accurate semantic mapping between subjects, as common subjects might be referred to with different phrases (or surface-forms) in different opinions.

Mapping opinion subjects: To map the different subjects among opinions, we first create a bipartite graph by computing a semantic similarity score between the subjects discussed in the opinions. This graph is then used to compute the mapping between the opinion subjects using Word Mover Distance (WMD) (Kusner et al. 2015), which aims at computing a minimum weight *perfect* matching between the opinion subjects. A detailed instantiation of this step in our framework is presented later section.

Aggregating polarity difference: Let S_i^1 and S_j^2 be the i^{th} and j^{th} opinion subjects in opinions O_1 and O_2, respectively. Also, let pol(S_i^k) represent the expressed polarity towards opinion subject i in opinion k (O_k). Then, the opinion distance between O_1 and O_2 is

$$OD(O_1, O_2) \;=\; \frac{\sum\limits_{(i,j)\in\mathcal{M}} f\,(\;pol(S_i^1)\;,\;pol(S_j^2)\;)}{2 \times |\mathcal{M}|} \qquad (1)$$

where $\mathcal{M}$ is the set of mapped opinion subjects. i and j are opinion subjects in O_1 and O_2, respectively and f is the difference function defined as

$$f(x,y) = \begin{cases} |x-y|, & \text{if x, y are absolute values} \\ JSD(x,y) \; or \; EMD(x,y), & \text{if x, y are distributions} \end{cases}$$

Here, JSD denotes Jenson-Shannon divergence and EMD denotes Earth Mover Distance. Both JSD and EMD are symmetric measures and EMD does not suffer from arbitrary quantization problems (Rubner, Tomasi, and Guibas 2000). Note, that the value of $OD(O_1, O_2)$ lies between [0,1].

Example: Consider an example with two opinions having opinion subjects as shown in bold below:
O_1: *"**Video games** increases violent tendencies among **youth**."*
O_2: *"**Researchers** have confirmed the potential positive effects of **computer games** and **media contents.**"*

We first compute the representation of the two opinions in terms of opinion subjects and polarity. So, O_1 is represented by the vector [("Video game", -1), ("youth", 0)] and O_2 is represented by [("researchers", 0), ("computer games", +1),

("media contents", +1)]. We then compute the semantic distance matrix which contains the semantic distance between the subjects in O_1 and O_2. Based on this matrix, our framework identifies that the subjects "video game" and "computer games" are highly similar, while the others are not. Let S_i^1 be the opinion subject "Video game", while S_j^2 be the opinion subject "Computer game". Further, the polarity of S_i^1, pol(S_i^1) is computed as "-1", while pol(S_j^2) is "+1"; and the difference function f(pol(S_i^1), pol(S_j^2)) is 2. Since this is the only mapping subject pair ($|\mathcal{M}|$=1), the final opinion distance OD(O_1, O_2) is obtained as 1 (max. value). Thus, the distance is large.

Opinion representations

In this section, we present two alternative representations for identifying opinion subjects. The first representation considers noun-phrases as opinion subjects and the dependent adjectives, adverbs, and verbs as associated words. The second representation consists of disambiguated concepts as opinion subjects with the words surrounding the subject as associated words. While the former relies on carefully defined rules to identify noun phrases and a careful analysis of the dependency parse tree to compute the polarity of the dependent associated words, the latter uses the weighted aggregation of the polarity of the neighboring words. We note that the latter is more efficient and avoids the computationally expensive and error-prone step of dependency parsing. We refer to the distance measures computed using the above variants as *OD-parse* and *OD*, respectively, and empirically compare their effectiveness in the experiments section.

Noun-phrase Representation

A parse tree represents the structure of a text string based on the syntax of the input language (in our case, the English language). A dependency parse tree captures the dependencies between different linguistic units in a sentence. We use Stanford CoreNLP (Manning et al. 2014) for part-of-speech (POS) tagging, coreference resolution and dependency parsing. Note, our framework is agnostic to such choices, and other tools like Google SyntaxNet (Andor et al. 2016) could also be used.

Opinion subjects extraction: In this approach, we consider all noun phrases in the opinion post as opinion subjects. We define a noun phrase based on some carefully defined rules and named-entities like Person, Organization, and Location. An exemplar of such a rule is,

$$(< \text{NN.*} >< \text{POS} >?) + < (\text{OF}|\text{THE}|\text{IN}) >? (< \text{NN.*} >< \text{POS} >?)) +$$

where NN denotes a noun word (based on Stanford CoreNLP POS tagging), POS denotes a possessive form, and the symbols '.*', '?', '+' are regex symbols as defined in Stanford CoreNLP. Observe, the above rule can capture even complex noun phrases like "J. R. R. Tolkien's Lord of the Rings".

Opinion expression extraction: To calculate the sentiment polarity expressed on the subject, we first identify all the related verbs, adverbs, and adjectives dependent on the noun phrase. For this, we use co-reference resolution and dependency parsing. For instance, Figure 2 shows the dependency parse tree of opinion O_1: *"Video game increases the violent tendencies among youth."*.

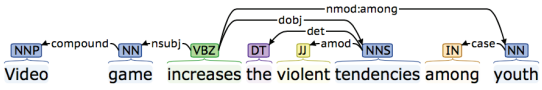

Figure 2: Dependency tree of sample sentence

Here, the noun phrase "Video game" would be extracted as an opinion subject while the verb "increases" is related and is part of how the subject "video game" is expressed. However, an efficient representation of opinions should also involve the terms "violent tendencies" in the opinion expression of "Video game". To extract the set of words for opinion expressions, we carefully defined 14 rules using Stanford CoreNLP's Semgrex pattern matching system.

Polarity of opinion subjects: We use the IBM Debator (Toledo-Ronen et al. 2018) sentiment lexicon to get the sentiment polarity score associated with individual words.[3] We aggregate the sentiment polarity scores of different words in the opinion expression in order to compute the sentiment expressed towards the corresponding subject. For this, we consider two techniques:

1. *Average and then discretize the polarity scores.* If the average sentiment polarity score of the words in the opinion expression of an opinion subject is negative, we assign a -1 score to the opinion subject, otherwise we assign it the score +1.

2. *Consider entire distribution of polarity scores.*

Overall, we found the parse-tree based approach to be computationally intensive. Furthermore, simply treating all noun phrases as opinion subjects resulted in the misidentification of many opinion subjects. For instance, in our example, "tendencies" would also act as an opinion subject. Another issue is that we do not capture any semantics, so synonyms get identified as different subjects while polysemies get identified as the same subject. Also, efficient extraction of opinion expressions requires a large number of rules and identifying all such rules is a manual and cumbersome process.

Disambiguated concept representation

In order to resolve the issues with the previous approach, we now describe an alternative approach.

Opinion subjects extraction: Ideally, the opinion subjects should be the key entities and concepts discussed in the opinion post. They should have a canonical representation, independent of the exact noun phrase used to discuss them, which can be achieved by leveraging *named-entity disambiguation* approaches.

In this work, we use the popular TagMe spotter to identify relevant noun phrases and the TagMe API (Ferragina

[3]Note that if an individual word has a negation modifier present in the dependency tree, then we multiply the sentiment polarity of that word by -1.

List of sentiment polarity shifters
'no', 'not', 'negation', 'none', "n't", 'inconclusive', 'without', 'excluded', "incompatible", "prevent", 'exacerbate', 'reduce', 'less', 'rarely', 'displaced', 'relocation', 'dislocation', 'higher than', 'relocate', 'resettled', 're-housed', 'cannot', 'limit', 'outweigh', 'unless', 'little act', 'get even'

Table 1: Sentiment polarity shifters used across all datasets.

and Scaiella 2010) for disambiguating the noun phrases to Wikipedia pages. The disambiguation process aids in understanding the noun phrases referring to the same entity/concept as well as those that are referring to entities/concepts that are semantically similar. For example, TagMe maps both the phrases "Video game" and "Electronic game" to the Wikipedia page of "Video game", while it maps "Computer game" to the semantically similar Wikipedia page titled "PC game".

Opinion expression extraction: To avoid the error-prone and computationally intensive process of fine-grained dependency parsing, we use all the adjacent sentiment words of an opinion subject S as the opinion expression towards S, similar to (Bar-Haim et al. 2017). However, not all adjacent words are treated equally. The influence of an opinion expression towards an opinion subject is weighted based on how far the opinion expression is from opinion subject in the sentence. Although in this relaxation many unrelated words may end up influencing the computation of sentiment polarity towards a subject, we show in the experiments section that this approach performs well in practice.

Polarity of opinion subjects: Similar to the previous approach, we use the IBM Debator (Toledo-Ronen et al. 2018) sentiment lexicon to find the sentiment polarity of words in an opinion expression. The sentiments of different words in the opinion expression are then aggregated by either considering the entire distribution or taking the average and then discretizing it. However in this case, before taking the average, we reweigh the sentiment value of a word by penalizing the sentiment score by a factor of $\sqrt{d}$ where d is the token distance of the word from opinion subject in the sentence. This ensures that the further a word is from the opinion subject, the less influence it has on its polarity score. Furthermore, if a polarity shifter (negator words like "no", "not" and "cannot") is present in the opinion expression, we reverse the polarity of the word. The complete list of polarity shifter words considered by our framework is reported in Table 1.

We then take the weighted average of distance-weighted sentiment scores (after polarity shifters). Hence, if p and n refer to the weighted sum of positive and negative sentiments associated with the subject, the weighted average is given by $(p-n)/(p+n+1)$. This is similar to (Feldman et al. 2011) and we empirically show that such a representation of opinion expression results in identifying better opinion clusters. Note that if an opinion subject is present in a sentence, all the words present in that sentence would influence the weighted computation of sentiment score of the subject. For example, our opinion subject "Video game" would now be represented as average weighted polarities of "increased", "violent", "tendencies", and "youth".

Interestingly, one could consider including the polarity of the opinion subject itself for computing the average polarity. However, this might incorrectly represent the opinion. For instance, consider the opinion "Genocide is good". Here, the opinion subject "Genocide" will have a positive polarity, correctly capturing what is expressed. Including the polarity of the subject, would lead to "Genocide" having a neutral or negative polarity – different from the expressed opinion.

Mappings between opinion subjects

To map the opinion subjects in the different opinions we create a bipartite graph by computing the semantic similarity score between the opinion subjects. Semantic similarity can be computed using embedding techniques such as word2vec (Mikolov et al. 2013) or doc2vec (Le and Mikolov 2014). In addition, the latter can use text similarity between corresponding Wikipedia page abstracts, number of common in-links and out-links between their Wikipedia pages, and so on. We then use the bipartite graph to compute the mapping between the opinion subjects using Word Mover Distance (WMD) (Kusner et al. 2015). An efficient linear time implementation of WMD exists (Pele and Werman 2008) to compute "the minimum distance that the embedded words of one document need to travel to reach the embedded words of another document". The underlying flow matrix in WMD computation provides a mapping between the subjects among the opinions.

WMD aims at computing a minimum weight *perfect* matching between the subjects. However, in practice, there may not be a good semantic mapping between all the subjects in the opinions, even if they are on the same topic.

Thus, the WMD mapping may also include subject pairs that have high semantic distance, resulting in erroneous comparisons. If two dissimilar subjects get mapped in WMD computation the overall distance between two increases. To address such erroneous mappings, we remove all pairs from the mapping whose semantic distance is greater than a user-defined threshold. If no mappings between two opinions exist after the semantic distance threshold, we set their opinion distance as undefined.

Experiments

Datasets and Empirical Setup

Datasets: One issue in the validation our proposed opinion distance is the lack of publicly available annotated ground truth datasets. However, there are many good benchmarks available for a restricted version of opinion clustering, namely stance detection. In stance detection, there are generally two labels – one supporting the topic-of-interest and the other opposing it. The assumption we make for evaluating opinion distances on the stance datasets is that "similar claims demonstrate similar stances (opinions)". We collect the stance detection datasets from IBM Debator Claims (Bar-Haim et al. 2017) and Arguments (Hou and Jochim 2017) projects. Table 2 provides a complete list of

Datasets
Seanad Abolition (25), Pornography (52), Gambling (60), Video Games (72), National Service (33), Monarchy (61), Hydroelectric Dams (110), Keystone pipeline (18), Democratization (76), Open-source Software (48), Intellectual Property (66), Atheism (116), Education Voucher Scheme (30), One-child policy China (67), Austerity Measures (20), Affirmative Action (81), Housing (30), Trades Unions (19)

Table 2: Dataset information. The number in brackets represents the number of opinions.

	Seanad Abolition		Video Games		Pornography	
Methods	ARI	Sil	ARI	Sil	ARI	Sil
TF-IDF	0.23	0.02	-0.01	0.01	-0.02	0.01
WMD	0.09	0.01	0.01	0.01	-0.02	0.01
Sent2vec	-0.01	-0.01	0.11	0.06	0.01	0.02
Doc2vec	-0.01	-0.03	-0.01	0.01	0.02	-0.01
BERT	0.03	-0.04	0.08	0.05	-0.01	0.03
OD-parse	0.01	-0.04	-0.01	0.02	0.07	0.05
OD	**0.54**	**0.31**	**0.56**	**0.42**	**0.41**	**0.41**

Table 3: ARI and Silhouette coefficient scores.

18 datasets used in our evaluation. We also evaluate our proposed methods against existing stance baselines (Pontiki and et al. 2016).

Additionally, we curate a nuanced opinion dataset from the civiq.eu platform discussing whether the Seanad (upper house of parliament in the Republic of Ireland) should be abolished. This dataset has three expert-curated opinion clusters (Liston 2013) (1) Abolish the Seanad, (2) Reform the Seanad instead of abolishing it, and (3) Seanad is ineffective but keep it until Dail is reformed to save the democracy. Note that the stance of the last opinion cluster is aligned with the second cluster – both against abolishing the Seanad. However, the general opinion expressed in the last cluster is to have the institution of Seanad only till the lower house of the parliament, Dail, is appropriately reformed. This is a nuanced argument not easily captured by existing methods.

Methods

In this section, we explain the selected baselines and the parameter setting for benchmarking the performance of the competing approaches.

TF-IDF: The distance between two opinions is computed as the cosine distance between their term-frequency inverse-document frequency (TF-IDF) vectors of those opinions (Schütze, Manning, and Raghavan 2008). Following standard practice, we remove the stop words while computing the TF-IDF.

WMD: The distance between two text documents can be captured using the recently proposed Word Mover Distance (WMD) (Kusner et al. 2015) – which in our setting would enable the capture of the distance between opinions. We use a pre-trained word2vec embedding (Mikolov et al. 2013) trained on GoogleNews Corpus for computing the WMD.

Sent2vec: The distance between two opinions is the cosine distance between Sent2vec (Pagliardini, Gupta, and Jaggi 2018) embeddings of those opinions. We lever the pretrain sent2vec-wiki-unigrams model for computing sentence embeddings.

Doc2vec: The distance between two opinions is the cosine distance between Doc2vec (Le and Mikolov 2014) embeddings of those opinions. We pre-train the Doc2vec methodology on Wikipedia articles.

BERT: The distance between two opinions is the cosine distance between BERT (Devlin et al. 2019) embeddings of those opinions. The embedding of opinions is computed by averaging the tokens' BERT embeddings using bert-as-a-service tool (Xiao 2018).

Our proposed methods are **OD-parse** and **OD**[4]. Unless noted otherwise, we use Doc2vec embeddings trained on Wikipedia articles with cosine distance for computing the semantic distance. The semantic distance threshold for our framework is set to 0.3. For TagMe, we select a link probability threshold of 3%.

Experimental results

We evaluate the methods in two different settings.

• **Unsupervised Setting**: We examine the utility of the proposed opinion distance measure in an unsupervised setting for opinion clustering. For the competing baselines, we compute the all-pair opinion distance matrix, and then evaluate the distance measure in this unsupervised setting based on clustering quality and on the intra- and inter- cluster distances. The clustering quality is measured using the Adjusted Rand Index (ARI), while evaluations based on intra and inter clusters distance is done with the Silhouette coefficient. We set the number of clusters equal to the number of unique opinion labels present in the dataset. We perform k-means clustering and spectral clustering and report the best ARI for all approaches. The quality of the identified opinion clusters (a snapshot) is shown in Table 3. We observe that:

The opinion distance measure captures the nuances among opinions: We see that *OD* significantly outperforms the baseline methods and the *OD-parse* variant[5]. The selected datasets contains set of contrasting opinions with same bag-of-words and as hypothesized earlier the existing text-similarity methods cannot understand the nuances among the opinions – visible from the low ARI and Sil scores. On the other hand, OD achieves high ARI and Sil scores, and seems to capture the nuances among opinions. Additionally, the noticeably high ARI and Silhouette coefficient values for OD seems to validate the observation of "similar opinions express similar sentiment polarity on discussed subjects".

Existing distance measures fail to capture nuances among opinions: From the above table, we observe that the

[4]The source code of our proposed method is available at https://github.com/saketguru/Opinion-Distance

[5]*OD-parse* fails to perform efficiently possibly due to poor opinion representation induced by erroneous opinion expression extraction based on the dependency parse tree.

Baselines	Seanad Abolition	Video Games	Pornography
Unigrams	0.54	0.66	0.63
Bigrams	0.54	0.64	0.56
LSA	0.68	0.57	0.57
Sentiment	0.35	0.60	0.69
Bigrams + Sentiment	0.43	0.58	0.66
TF-IDF	0.50	0.65	0.57
WMD	0.40	0.73	0.57
Sent2vec	0.39	0.79	0.70
Doc2vec	0.27	0.51	0.56
BERT	0.46	0.84	0.68
Unigrams			
+ Bigrams	0.40	0.64	0.78
+ Sentiment	0.24	0.54	0.54
+ LSA	0.73	0.51	0.58
+ TF-IDF	0.42	0.65	0.56
+ WMD	0.48	0.73	0.53
+ Sent2vec	0.56	0.59	0.66
+ Doc2vec	0.31	0.56	0.47
OD-parse	0.50	0.58	0.53
OD	0.71	**0.88**	0.88
OD			
+ Unigrams	0.83	*0.86*	*0.88*
+ Bigrams	**0.87**	0.85	*0.88*
+ Sentiment	0.64	*0.86*	0.86
+ LSA	*0.84*	0.82	**0.90**
+ WMD	0.75	0.82	0.86

Table 4: The quality of opinion distance when leveraged as a feature for multi-class classification. Each entry in + *X* feature should be treated independently. The second best result is italicized and underlined.

text-similarity based baselines, such as TF-IDF, WMD and Doc2vec achieving ARI and Silhouette coefficient scores of close to zero on the "Video Games" and "Pornography" datasets (barely providing a performance improvement over random clustering, i.e., a zero ARI score).

A possible source for the poor performance might be the underlying assumption that similar opinions should have common words and/or a similar word distribution – which does not hold in cases of opinion as depicted in previous sections. A notable exception is the "Seanad Abolition" dataset, where TF-IDF performs relatively better than WMD, Sent2vec and Doc2vec. Drilling down we find that a highly discriminatory word "democracy" occurs in only one of the clusters (namely "Save the Democracy") - explaining its performance on this dataset.

• **Supervised Setting**: We now examine the effectiveness of our approach in a supervised setting. Here, we leverage the idea that distance or similarity based problems can be reformulated as standard classification problems by treating pairwise similarities as features to a downstream classification method like SVM (Graepel et al. 1999; Pekalska, Paclik, and Duin 2001; Liao and Noble 2003). Specifically, for all the approaches, we treat the distance measure as a feature by considering each row i of the distance matrix as a feature vector for opinion i and the task is to check whether two opinions have same label or not. For all the classification experiments, unless otherwise noted, we use SVM as classifier with RBF kernel.[6] We perform hyperparameter tuning using both grid search (with 5-fold cross validation) and auto-sklearn (Feurer et al. 2015) (with default 'holdout' strategy). We use two tuning strategies because we face the problem of overfitting for smaller datasets using only one of the tuning strategy. Since we use two tuning strategies, the hyperparameter tuning is done on train split (70%) and we report the best results on test split (30%) averaged over 3 runs. For evaluation, we rely on the average weighted F1 measure for classification accuracy. For completeness, here we also compare against unigram or n-gram based classifiers that typically work well in such settings (Pontiki and et al. 2016; Somasundaran and Wiebe 2010). The classification performance of the baselines is reported in Table 4. We observe that:

	Difference Function	Seanad Abolition	Video Games	Pornography
OD-parse	Absolute	0.01	-0.01	0.07
	JS div.	0.01	-0.01	-0.01
	EMD	0.07	0.01	-0.01
OD	Absolute	**0.54**	**0.56**	**0.41**
	JS div.	0.07	-0.01	-0.02
	EMD	0.26	-0.01	0.01
OD (no polarity shifters)	Absolute	0.23	0.08	0.04
	JS div.	0.09	-0.01	-0.02
	EMD	0.10	0.01	-0.01

Table 5: We compare the quality of variants of Opinion Distance measures on opinion clustering task with ARI.

SVM with only OD features outperforms many baselines: We see that on "Video Games" and "Pornography" datasets, the classification performance based on SVM with only OD is significantly better than the SVM with any other combination of features excluding OD. For the "Seanad Abolition" dataset, there is one exception: SVM with unigrams and LSA features performs slightly better than OD. As discussed earlier, this can be attributed to the discriminating word "democracy".

SVM with OD and baseline features further improves classification performance: We see that SVM with OD and bigrams achieves the best multi-class classification performance on the "Seanad Abolition" dataset. On the "Pornography" dataset, we observe SVM with OD + LSA to improve classification performance by nearly 2%.

Opinion Distance Drilldown

In this section, we present a detailed drilldown of our proposed opinion distance measure. We consider the following two variants of our measure: *OD-d2v* and *OD-w2v* where the semantic distance is computed as cosine distance over Doc2vec embeddings (pre-trained on Wikipedia) and Word2vec embeddings (pre-trained on Google) respectively.

[6]Note that, we achieve similar score for the distance measure with Logistic Regression classifier.

Topic Name	Size	TF-IDF ARI	WMD ARI	Sent2vec ARI	Doc2vec ARI	BERT ARI	OD-w2v ARI	OD-d2v ARI	TF-IDF *Sil.*	WMD *Sil.*	Sent2vec *Sil.*	Doc2vec *Sil.*	BERT *Sil.*	OD-w2v *Sil.*	OD-d2v *Sil.*
Affirmative Action	81	-0.07	-0.02	0.03	-0.01	-0.02	**0.14**	*0.02*	0.01	0.01	-0.01	-0.02	-0.04	**0.06**	*0.01*
Atheism	116	**0.19**	0.07	0.00	0.03	-0.01	0.11	*0.16*	0.02	0.01	0.02	0.01	0.01	*0.05*	**0.07**
Austerity Measures	20	*0.04*	*0.04*	-0.01	-0.05	0.04	**0.21**	-0.01	0.06	0.07	0.05	-0.03	0.10	**0.19**	0.1
Democratization	76	0.02	-0.01	0.00	*0.09*	-0.01	**0.11**	0.07	0.01	0.01	0.02	0.02	0.03	**0.16**	*0.11*
Education Voucher Scheme	30	**0.25**	0.12	0.08	-0.02	0.04	0.13	*0.19*	0.01	0.01	0.01	-0.01	0.02	*0.38*	**0.40**
Gambling	60	-0.06	-0.01	-0.02	0.04	0.09	*0.35*	**0.39**	0.01	0.02	0.03	0.01	0.09	**0.30**	*0.22*
Housing	30	0.01	-0.01	-0.01	-0.02	0.08	**0.27**	0.01	0.02	0.03	0.03	0.01	0.11	**0.13**	*0.13*
Hydroelectric Dams	110	**0.47**	*0.45*	*0.45*	-0.01	0.38	0.35	0.14	0.04	0.08	0.12	0.01	0.19	**0.26**	*0.09*
Intellectual Property	66	0.01	0.01	0.00	0.03	0.03	*0.05*	**0.14**	0.01	*0.04*	0.03	0.01	0.03	*0.04*	**0.12**
Keystone pipeline	18	0.01	0.01	0.00	-0.13	**0.07**	-0.01	**0.07**	-0.01	-0.03	-0.03	-0.07	0.03	**0.05**	*0.02*
Monarchy	61	-0.04	0.01	0.00	0.03	-0.02	**0.15**	**0.15**	0.01	0.02	0.02	0.01	0.01	**0.11**	*0.09*
National Service	33	0.14	-0.03	-0.01	0.02	0.01	*0.31*	**0.39**	0.02	0.04	0.02	0.01	0.02	**0.25**	**0.25**
One-child policy China	67	-0.05	0.01	**0.11**	-0.02	0.02	**0.11**	0.01	0.01	0.02	*0.04*	-0.01	0.03	**0.07**	-0.02
Open-source Software	48	-0.02	-0.01	*0.05*	0.01	0.12	**0.09**	-0.02	0.01	-0.01	0.00	-0.02	0.03	**0.18**	0.01
Pornography	52	-0.02	0.01	0.01	-0.02	-0.01	**0.41**	**0.41**	0.01	0.01	0.02	-0.01	0.03	**0.47**	*0.41*
Seanad Abolition	25	0.23	0.09	-0.01	-0.01	0.03	*0.32*	**0.54**	0.02	0.01	-0.01	-0.03	-0.04	*0.15*	**0.31**
Trades Unions	19	*0.44*	*0.44*	**0.60**	-0.05	0.44	*0.44*	0.29	0.1	0.17	0.21	0.01	0.26	**0.48**	*0.32*
Video Games	72	-0.01	0.01	0.12	0.01	0.08	*0.40*	**0.56**	0.01	0.01	0.06	0.01	0.05	*0.32*	**0.42**
Average	54.67	0.09	0.07	0.08	0.01	0.08	**0.22**	*0.20*	0.02	0.03	0.04	-0.01	0.05	**0.20**	*0.17*

Table 6: Performance comparison of the distance measures on all 18 datasets. The semantic distance in opinion distance (OD) measure is computed via cosine distance over either Word2vec (OD-w2v with semantic distance threshold 0.6) or Doc2vec (OD-d2v with distance threshold 0.3) embeddings. $Sil.$ refers to Silhouette Coefficient. The second best result is italicized and underlined. The ARI and Silhouette coefficients scores of both *OD* methods (*OD-d2v* and *OD-w2v*) are statistically significant (paired t-test) with respect to baselines at significance level 0.005.

The semantic threshold for OD-d2v is set at 0.3 while for OD-w2v is set at 0.6. In both the variants, we use the sentiment polarity shifters, a complete list of which are presented in Table 1. We evaluate our distance measures in the unsupervised setting, specifically, evaluating the clustering quality using the Adjusted Rand Index (ARI) and Silhouette coefficient. We benchmark against the following baselines: WMD (which relies on word2vec embeddings), Doc2vec and TF-IDF. The results are shown in Table 6. The ARI and Silhouette coefficients scores of both *OD* methods (*OD-d2v* and *OD-w2v*) are statistically significant (paired t-test) with respect to baselines at significance level 0.005. We observe the following trends:

1. Opinion distance methods generally outperform the competition on both ARI and Silhouette coefficient. We observe that given a topic, the opinion distance measure is able to separate pro stance opinions from the con stance opinions. We also find that the other baselines generally perform worse as both pro and con stance opinions have high text similarity in many of these datasets. This is reflected in the average ARI and average Silhouette coefficients of the baseline distance measures.

2. On a few datasets, there are a few discriminating words between the different opinion clusters. For instance, in the topic "Trade Union", the opinion subject "collective bargaining" is contained in 60 % of claims in the con stance of the topic while it is not present in the pro stance opinions on the topic. The pro stance opinions on this topic use the term "Unions" as the opinion subject. Another example is "Hydroelectric Dams" dataset, where 36 % of pro stance opinions contain term "hydro" while only 18% of con stance opinions contain term "hydro". On such datasets, TF-IDF and WMD perform relatively better in separating pro stance opinions from the con stance opinions. In the exceptional case of "Hydroelectric Dams" dataset, the opinion distance OD performs particularly bad compared to TF-IDF because of some errors with disambiguation.

Variants of opinion distance

Next, we compare the different variants of opinion distance. As mentioned earlier, the polarity of opinion subject $pol(S_i)$ can be represented in the form of absolute value or polarity distribution. The results of different variants are shown in Table 5. We make the following observations.

OD **significantly outperforms** OD**-**$parse$**:** We observe that compared to OD-$parse$, OD is much more accurate. On the three datasets, OD achieves ARI score of 0.54, 0.56 and 0.41 respectively compared to the scores of 0.01, -0.01 and 0.07 by OD-$parse$. This is largely because of the errors in dependency parsing and the resultant poor opinion representation.

Representing polarities in form of distribution is inefficient: We find that irrespective of the chosen variant of opinion distance, if we represent $pol(S_i)$ in the form of distribution, we do not find good opinion clusters. The D value of opinion distance calculated using polarity distribution is also low. Recall that Jensen-Shannon divergence is the distance measure used in (Fang et al. 2012). Table 5 shows that OD is significantly better than Jensen-Shannon divergence on all the three datasets. **Sentiment polarity shifters have a high impact on clustering performance of opinion distance:** We find that not utilizing the sentiment polarity shifters, especially in case of datasets "Video games" and "Pornography" hurts the *Opinion Representation* phase, and thereby leads to incorrect computation of opinion distance. This is evident from the significant drop in ARI score from OD to OD (no polarity shifters) since the only change in those variants is of sentiment polarity shifters.

Case study

Here we describe a case study for navigating opinions expressed in the form of comments on a BBC news article

Title: Brexit: Second Commons defeat for Theresa May in 24 hours
O_1: "What's most detestable is the way people on here protest about how Brexit is an almost divine right of the people which should be delivered as though a birthright or destiny. It's not. It's really not. It was a snapshot of public opinion at a particular time point which was influenced by a lot of spin and subterfuge. Now is a different time. Time for this horror to end."
Below opinions O_2 and O_3 are similar to O_1
O_2 : "Not all leavers are thugs, i know because i voted leave and i'm not a thug. But there's a lot of leavers that have their head in the sand and can't admit that brexit is so complex that they and i didn't really understand the consequences. I'm not one of them leavers either as i've listened to the facts now and changed my mind. New factual vote please."
O_3: "STOP BREXIT, SAVE BRITAIN! They should jst go straight on to #revokeA50. It would save lots of wasted time, money and even the economy. This is now crunch time for Brexit. It looks like the doors have been closed for a No Deal exit and Theresa May's plan will most likely be voted down. The only remaining option therefore is to call the whole thing off and let us get back to normal. "

Table 7: Similar comments identified with the help of the proposed OD measure on (BBC.com 2019).

Title: Theresa May urges Jeremy Corbyn: Let's talk Brexit
O_1: "The last thing Comrade Corbyn wants is cross-party consensus on Brexit. That would get in the way of his desperate plan to force a general election. It is now obvious to everyone (I hope) that this horrible man will say and do anything to get into power. He belongs in a glass case in a political museum, spouting his far-left communist trash, a bit like one of those old laughing sailor machines."
O_2: "The longer this goes on, the less it has to do with the EU. Brexit has become a conflict between those who value truth and individual rights above all else, and those for whom being in a majority (real, "silent" or based on a flawed referendum) is their identity. That's why the latter are so perpetually angry: they got what they wanted in 2016 but will lose it if facts win over fantasy in the end."

Table 8: Two comments expressing opposite opinions; OD correctly identifies them as dissimilar while Doc2vec considers them very similar.

	SemRegex Rules
1	$\{\} = OpExp1 > /nmod:in/\{\} = OpSubject$
2	$\{\} = OpExp1 > /nmod:to/\{\} = OpSubject$
3	$\{\} = OpExp1 < /nmod:of/\{\} = OpSubject$
4	$\{\} = OpExp1 > /nmod:by/\{\} = OpSubject$
5	$\{\} = OpExp1 > /nmod:after/\{\} = OpSubject$
6	$\{\} = OpExp1 > /nmod:without/\{\} = OpSubject$
7	$\{\} = OpExp1 > nsubj\{\} = OpSubject > dobj\{\} = OpExp2$
8	$\{\} = OpExp1 > nsubj\{\} = OpSubject < csubj\{\} = OpExp2$
9	$\{\} = OpExp1 > nsubj\{\} = OpSubject > /compound.*/\{\} = OpExp2$
10	$\{\} = OpExp1 < /advcl:as/\{\} = OpSubject$
11	$\{tag:/VB.*/\} > nsubj\{\} = OpSubject [> acomp\{\} = OpExp1 \mid > xcomp\{\} = OpExp2]$
12	$\{tag:/VB.*/\} = OpExp2 > advmod \{\} = OpExp1[< \{\} = OpSubject \mid > \{\} = OpSubject]$
13	$\{tag:/NN.*/\} > nsubj\{\} = OpSubject > amod\{\} = OpExp1$
14	$\{tag:/NN.*/\} > nsubj\{\} = OpSubject > /nmod.*/(\{\} = OpExp1 > amod\{\} = OpExp2$

Table 9: SemRegex Rules

"Brexit: Second Commons defeat for Theresa May in 24 hours" (BBC.com 2019). Table 7 shows an example of comments that were correctly identified to be similar by OD, but reported highly dissimilar by TF-IDF and WMD. These comments are clearly against Brexit and want it to either be called off or be subjected to another referendum. While in the above example the Doc2vec distance was also smaller, it fails to capture the dissimilarity between comments that are very different. For instance, Table 8 presents one comment that is highly critical of Labour leader Jeremy Corbyn, while the second comment opines poorly on the ruling conservative party; Doc2vec identifies them as similar while OD correctly identifies them as highly dissimilar.

Conclusion

Automated tools for better understanding of opinions would lead us to a new era of digital democracy and improved decision making. In this work, we proposed an opinion distance measure for quantifying the distance between opinions. Based on the observation that similar opinions express similar sentiment polarity on discussed subjects, we show that our proposed measure significantly outperforms existing approaches in both unsupervised and supervised experimental setups. We believe the progress in the research of nuanced opinion analysis is hindered by the lack of publically available large opinion datasets. In the future, we plan to work on the construction of large opinion datasets con-

sisting of multiple nuanced opinion clusters.

Acknowledgements

This material is based upon work supported by the National Institute of Health (NIH) under Grant No NIH-1R01 HD088545-01A1 and National Science Foundation (NSF) under Grant No EAR-1520870. Computational support was provided by the Ohio Supercomputer Center under grant PAA0205. Any opinions, findings, and conclusions in this material are those of the author(s) and may not reflect the views of the respective funding agency.

Additional Details About Noun-Phrase Representation: SemRegex Rules

The parse-tree variant of opinion distance $OD-parse$ relies on the dependency parse tree to extract opinion subject and opinion expression terms. In addition to the basic dependencies, enhanced dependencies and Named entity extraction functionalities present in Stanford CoreNLP, we also lever SemRegex rules. The list of rules with Stanford CoreNLP's SemRegrex pattern matching system is presented in Table 9. The $OpSubject$ refers to opinion subject while $OpExp$# refers to the opinion expression.

References

Andor, D.; Alberti, C.; Weiss, D.; Severyn, A.; Presta, A.; Ganchev, K.; Petrov, S.; and Collins, M. 2016. Globally normalized transition-based neural networks. In *ACL*.

Awadallah, R.; Ramanath, M.; and Weikum, G. 2012. Harmony and dissonance: organizing the people's voices on political controversies. In *WSDM*.

Bar-Haim, R.; Bhattacharya, I.; Dinuzzo, F.; Saha, A.; and Slonim, N. 2017. Stance classification of context-dependent claims. In *EACL*.

2019. BBC News. https://www.bbc.com/news/uk-politics-46805269.

Buche, A.; Chandak, D.; and Zadgaonkar, A. 2013. Opinion mining and analysis: a survey. *arXiv preprint arXiv:1307.3336*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *ACL*.

Fang, Y.; Si, L.; Somasundaram, N.; and Yu, Z. 2012. Mining contrastive opinions on political texts using cross-perspective topic model. In *WSDM*, 63–72.

Feldman, R.; Rosenfeld, B.; Bar-Haim, R.; and Fresko, M. 2011. The stock sonar sentiment analysis of stocks based on a hybrid approach. In *IAAI*.

Ferragina, P., and Scaiella, U. 2010. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *CIKM*.

Feurer, M.; Klein, A.; Eggensperger, K.; Springenberg, J.; Blum, M.; and Hutter, F. 2015. Efficient and robust automated machine learning. In *NeurIPS*.

Graepel, T.; Herbrich, R.; Bollmann-Sdorra, P.; and Obermayer, K. 1999. Classification on pairwise proximity data. In *NeurIPS*, 438–444.

Hou, Y., and Jochim, C. 2017. Argument relation classification using a joint inference model. In *Workshop on Argument Mining*.

Kim, S.-M., and Hovy, E. 2006. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Workshop on Sentiment and Subjectivity in Text, ACL*.

Kobayashi, N.; Inui, K.; Matsumoto, Y.; Tateishi, K.; and Fukushima, T. 2004. Collecting evaluative expressions for opinion extraction. In *International Conference on Natural Language Processing*, 596–605. Springer.

Kusner, M.; Sun, Y.; Kolkin, N.; and Weinberger, K. 2015. From word embeddings to document distances. In *ICML*.

Le, Q., and Mikolov, T. 2014. Distributed representations of sentences and documents. In *ICML*, 1188–1196.

Liao, L., and Noble, W. S. 2003. Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. In *Journal of computational biology*.

Liston, V. 2013. Perspective on seanad abolition. https://politicalreform.ie/2013/10/11/perspective-on-seanad-abolition/.

Liu, B., and Zhang, L. 2012. A survey of opinion mining and sentiment analysis. In *Mining text data*.

Liu, B.; Hu, M.; and Cheng, J. 2005. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, 342–351. ACM.

Manning, C.; Surdeanu, M.; Bauer, J.; Finkel, J.; Bethard, S.; and McClosky, D. 2014. The stanford corenlp natural language processing toolkit. In *ACL*.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. In *arXiv preprint arXiv:1301.3781*.

Mohammad, S.; Kiritchenko, S.; Sobhani, P.; Zhu, X.; and Cherry, C. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *SemEval*.

Mullen, T., and Collier, N. 2004. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of the 2004 conference on empirical methods in natural language processing*.

Pagliardini, M.; Gupta, P.; and Jaggi, M. 2018. Unsupervised learning of sentence embeddings using compositional n-gram features. In *NAACL*.

Pang, B.; Lee, L.; et al. 2008. Opinion mining and sentiment analysis. In *Foundations and Trends® in Information Retrieval*.

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *ACL*, 79–86. Association for Computational Linguistics.

Pavlopoulos, I. 2014. Aspect based sentiment analysis. *Athens University of Economics and Business*.

Pekalska, E.; Paclik, P.; and Duin, R. P. 2001. A generalized kernel approach to dissimilarity-based classification. In *JMLR*.

Pele, O., and Werman, M. 2008. A linear time histogram metric for improved sift matching. In *ECCV*.

Pontiki, M., and et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *SemEval*.

Quraishi, M.; Fafalios, P.; and Herder, E. 2018. Viewpoint discovery and understanding in social networks. In *WebSci*.

Recupero, D.; Presutti, V.; Consoli, S.; Gangemi, A.; and Nuzzolese, A. 2015. Sentilo: frame-based sentiment analysis. In *Cognitive Computation*.

Riedel, B.; Augenstein, I.; Spithourakis, G. P.; and Riedel, S. 2017. A simple but tough-to-beat baseline for the fake news challenge stance detection task. In *arXiv preprint arXiv:1707.03264*.

Rubner, Y.; Tomasi, C.; and Guibas, L. J. 2000. The earth mover's distance as a metric for image retrieval. *International journal of computer vision* 40(2):99–121.

Schütze, H.; Manning, C. D.; and Raghavan, P. 2008. *Introduction to information retrieval*. Cambridge University Press.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A.; and Potts, C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*.

Somasundaran, S., and Wiebe, J. 2010. Recognizing stances in ideological on-line debates. In *NAACL*. Association for Computational Linguistics.

Stab, C., and et al. 2018. Argumentext: Searching for arguments in heterogeneous sources. In *ACL*.

Titov, I., and McDonald, R. 2008. A joint model of text and aspect ratings for sentiment summarization. *ACL*.

Toledo-Ronen, O.; Bar-Haim, R.; Halfon, A.; Jochim, C.; Menczel, A.; Aharonov, R.; and Slonim, N. 2018. Learning sentiment composition from sentiment lexicons. In *COLING*.

Trabelsi, A., and Zaïane, O. R. 2018. Unsupervised model for topic viewpoint discovery in online debates leveraging author interactions. In *ICWSM*.

Wiegand, M.; Schulder, M.; and Ruppenhofer, J. 2015. Opinion holder and target extraction for verb-based opinion predicates – the problem is not solved. In *Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.

Xiao, H. 2018. bert-as-service. https://github.com/hanxiao/bert-as-service.

Zhang, L.; Wang, S.; and Liu, B. 2018. Deep learning for sentiment analysis: A survey. In *Data Mining and Knowledge Discovery*.