Undefined 1 (2009) 1–5
IOS Press

Sentiment Lexicon Adaptation with Context and Semantics for the Social Web

Editor(s): Name Surname, University, Country
Solicited review(s): Name Surname, University, Country
Open review(s): Name Surname, University, Country

Hassan Saif [a,*], Miriam Fernandez [a], Leon Kastler [b] and Harith Alani [a]
[a] *Knowledge Media Institute, Open University, MK76AA, Milton Keynes, UK*
E-mail: {h.saif,m.fernandez,h.alani}@open.ac.uk
[b] *Institute for Web Science and Technology, University of Koblenz-Landau, 56070 Koblenz, Germany*
E-mail: lkastler@uni-koblenz.de

Abstract. Sentiment analysis over social streams offers governments and organisations a fast and effective way to monitor the publics' feelings towards policies, brands, business, etc. General purpose sentiment lexicons have been used to compute sentiment from social streams, since they are simple and effective. They calculate the overall sentiment of texts by using a general collection of words, with predetermined sentiment orientation and strength. However, words' sentiment often vary with the contexts in which they appear, and new words might be encountered that are not covered by the lexicon, particularly in social media environments where content emerges and changes rapidly and constantly. In this paper, we propose a *lexicon adaptation* approach that uses contextual as well as semantic information extracted from DBPedia to update the words' weighted sentiment orientations and to add new words to the lexicon. We evaluate our approach on three different Twitter datasets, and show that enriching the lexicon with contextual and semantic information improves sentiment computation by 3.7% in average accuracy, and by 3% in average F1 measure.

Keywords: Sentiment Lexicon Adaptation, Semantics, Twitter

1. Introduction

Sentiment analysis on social media, and particularly on Twitter, has gained much attention in recent years. Twitter offers a platform where users often express their opinions and attitudes towards a great variety of topics, offering governments and organisations a fast and effective way to monitor the publics' feelings towards their brand, business, policies, etc.

However, sentiment analysis over social media data poses new challenges due to the typical ill-formed syntactical and grammatical structures of such content [26]. Although different type of approaches have been proposed in the last few years to extract sentiment over this type of data, Lexicon-based approaches have gained popularity because, as opposed to Machine Learning approaches, they do not require the use of training data, which is often expensive and/or impractical to obtain. These approaches use general-purpose sentiment lexicons (sets of words with associated sentiment scores) to compute the sentiment of a text regardless of its domain or context [4,18,29,11]. However, a word's sentiment may vary according to the context in which the word is used [31]. For example, the word *great* conveys different sentiment when associated with the word *problem* than with the word *smile*. Therefore, the performance of these lexicons may drop when used to analyse sentiment over specific domains or contexts.

Some works have attempted to address this problem by generating domain-specific lexicons from scratch [3,8,16,14], which tends to be costly, especially when applied to dynamic and generic microblog data (e.g.,

*Corresponding author. E-mail: hassan.saif@open.ac.uk.

[6,9]). Others opted for extending popular lexicons to fit new domains [7,13,27,12]. Automatic adaptation of existing lexicons not only reduces the burden of creating a new lexicon, but also ensures that the words' sentiment and weights generated and tested during the construction of existing lexicons are taken into consideration as basis for adaptation [7,20].

In addition, very little attention has been giving to the use of semantic information as a resource to perform such adaptations. Our hypothesis is that semantics can help to better capture the domain or context for which the lexicon is being adapted, thus aiming to contribute towards a more informed calculation of words' sentiment weights. For example, the context of the word "`Ebola`" in "`Ebola continues spreading in Africa!`" does not indicate a clear sentiment for the word. However, "`Ebola`" is associated with the semantic type (concept) "`Virus/Disease`", which suggests that the sentiment of "`Ebola`" is likely to be negative.

In this paper, we propose a general method to adapt sentiment lexicons to any given domain or context, where *context is defined by a collection of microblogs (Tweets)*. A key novelty of our method is that it not only captures the domain context (*contextual or distributional semantics*), it also introduces the use of *conceptual semantics*, i.e., semantics extracted from background ontologies such as DBpedia. In performing our study we make the following contributions:

1. We introduce a generic, unsupervised, method for adapting existing sentiment lexicons to given domains and contexts, defined by a collection of microblog posts (Tweets)
2. We propose two methods for semantically-enriching the lexicon adaptation method: (i) enrichment with the semantic concepts of words, and (ii) enrichment based on the semantic relations between words in tweets
3. We study three lexicon adaptation techniques: updating the words' sentiment weights, expanding the lexicon with new words, and the combination of both
4. We evaluate our context-based lexicon adaptation method over three Twitter datasets, and show an average, statistically significant, improvement of 3.4% in accuracy, and 2.8% in F1, against the baselines methods
5. We investigate the impact of the proposed semantic-enrichment approaches on the lexicon adaptation performance, and show that enrichment with words' semantic concepts, when used for updating the the words' sentiment weights in the lexicon, increases performance slightly over the context-based method by 0.27% and 0.25% in accuracy and F1 respectively. Enrichment based on the semantic relations between entities in tweets, yields in 4.12% and 3.12% gain in accuracy and F1, when used for expanding the lexicon with new opinionated words, in comparison with lexicon expanding without semantic enrichment
6. We investigate the impact of dataset imbalance when using lexicons for calculating tweet-level sentiment and show that our adapted lexicons have higher tolerance to imbalanced datasets

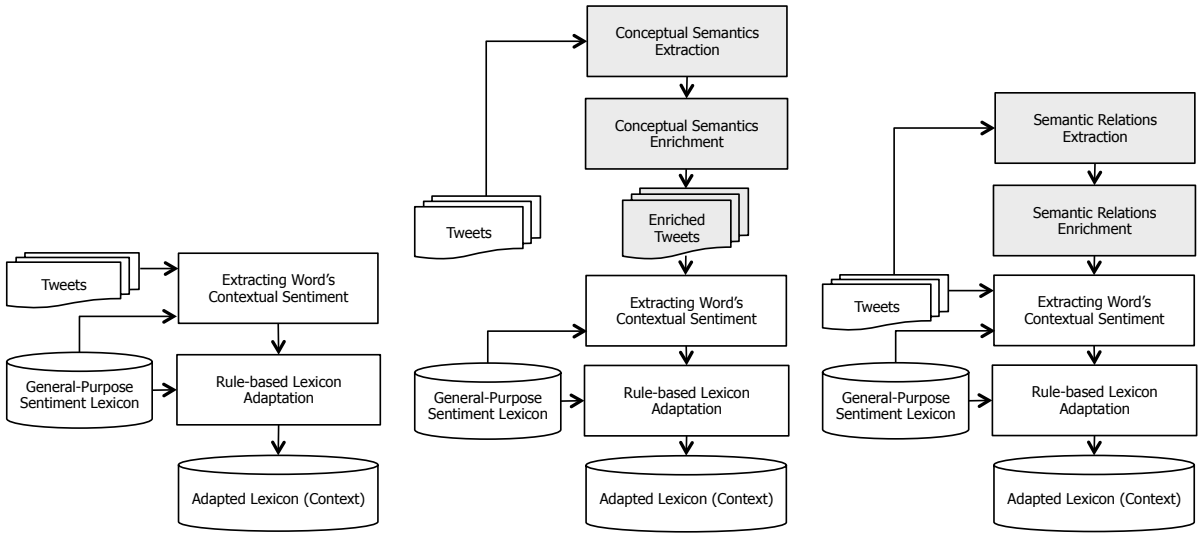
The remainder of this paper is structured as follows. Related work is discussed in Section 2. Our method for sentiment lexicon adaptation and its semantic enrichment is presented in Sections 3 and 4 respectively. Experimental setup and results are in Sections 5 and 6 respectively. Section 7 covers discussion and future work. Conclusions are reported in Section 8.

2. Related Work

General purpose sentiment lexicons (MPQA[18], SentiWordNet[2], Thelwall-lexicon[30], Nielsen-Lexicon[18]) have been traditionally used in the literature to determine the overall sentiment of texts. These lexicons capture a selection of popular words and their associated weighted sentiment orientations, without considering the domain, topic, or context where the lexicons are being used. However, a word's associated sentiment may vary according to the context in which the word is used [31]. To address this problem multiple works have emerged in recent years to: (i) create domain-specific lexicons or, (ii) adapt existing lexicons to specific domains.

Most existing works belong to the first category, where approaches have been proposed to develop sentiment lexicons tailored for specific domains [3,8,13, 16,27,14]. Works like [3,14,27] propose the use of bootstrapping methods for building sentiment lexicons. These methods use seed sets of subjective words [3,27], dictionaries [3,16], domain-specific corpora [3,13,8,16], training data from related domains [14], ratings [16] and graph-based formalisms [27] to identify words and to induce sentiment weights.

Recently, several works were focused on the development of domain-specific sentiment lexicons for social media [6,9]. To infer words and sentiment weights



(a) Context-based Adaptation (b) Conceptually-enriched Adaptation (c) Adaptation with Semantic Relation Adjustment

Fig. 1. Pipelines for (a) Context-based Adaptation Model, (b) Context-based Adaptation Model with Semantic Concepts Enrichment, and (c) Context-based Adaptation Model with Semantic Relations Adjustment

these approaches make use of linguistic and statistical features from tweets [9], including emoticons [12]. It is important to highlight that the informality of the language used in this type of data makes building domain-specific sentiment lexicons a more difficult task.

Rather than creating domain-specific sentiment lexicons, several approaches have proposed methods for adapting existing, well-known lexicons, to specific domains [7,20,24]. As previously mentioned, lexicon adaptation not only reduces the burden of creating lexicons from scratch, but also supplements the process with a collection of pre-existing words and their sentiment orientation and weights.

While the majority of work on lexicon adaptation focuses on conventional text, lexicon adaptation for social media data is still in its infancy. One very recent work in this line [12] has focused on updating the sentiment of neutral words in SentiWordNet. In addition to this work, we not only adapt sentiment weights, but also study the extraction and addition of new terms not provided in the original lexicon[24]. This is potentially useful in the case of social media data, where new terms and abbreviations constantly emerge. Note that, in-line with the work of Lu and colleagues [16], our proposed lexicon adaptation method is not restricted to domain-adaptation, but rather considers a more fine-grained context adaptation, where the context is defined by a collection of posts. Moreover, our approach does not make use of training data to adapt the lexicon.

Another novelty of our approach with respect to previous works, is the use of conceptual semantics, i.e. semantics extracted from ontologies such as DBPedia, to adapt sentiment lexicons. Our hypothesis is that conceptual semantics can help to better capture the domain for which the lexicon is being adapted, by enabling the discovery of relevant concepts and semantic relations between terms. The extraction of these concepts and relations (e.g., knowing that "`Ebola`" is an entity appearing within the tweet collection, and that it is associated with the semantic type (concept) "`Virus/Disease`") facilitates a better enrichment of the context and provides a higher term relationship coverage for the calculation of sentiment weights. Capturing the relationships among terms helps inferring the sentiment influence that terms have on one another within the context.

3. Context-based Lexicon Adaptation

The main principle behind lexicon adaptation is that the sentiment of a term is not as static as given in general-purpose sentiment lexicons, but it rather depends on the context in which the term is used [25]. In this section we present our method for adapting sentiment lexicons based on words' context in tweets.

The pipeline of our proposed context-based lexicon adaptation method consists of two main steps, as depicted in Figure1(a). First, given a tweet collection and a general-purpose sentiment lexicon, our approach detects the context of each word in the tweet collection and uses it to extract the word's contextual sentiment.

Secondly, a set of rules are applied to amend the prior sentiment of terms in the lexicon based on their corresponding contextual sentiment. Both steps are further detailed in Sections 3.1 and 3.2. The semantic enrichment of this pipeline is described in Section 4. Conceptual semantics are used to enrich the context or domain in which the words are used, with the aim of enabling a better interpretation of this context.

3.1. Word's Contextual Sentiment

The first step in our pipeline is to extract the contextual sentiment of terms (i.e., sentiment extracted based on a word's context) in a given tweet collection. This step consists of: (i) capturing the context in which the word occurs, and (ii) computing the word's contextual sentiment. A common method for capturing the word's context is by looking at its co-occurrence patterns with other terms in the text. The underlying principle behind this method comes from the distributional semantic hypothesis:[1] *words that are used and occur in the same contexts tend to purport similar meanings* [10,33]. For example, the word "great", when occurs in the context "smile", denotes a different meaning than when it occurs within the context "pain" and "loss". Such context variations of the word often affect its sentiment: "great" with "smile" indicates a positive sentiment while "great" with "pain" indicates a negative one.

Several approaches have been built and used for extracting the words' contextual sentiment following the above principle [32,15]. In this paper, we use the SentiCircle approach [22], which similarly to other frequency-based approaches, it detects the context of a term from its co-occurrence patterns with other terms in tweets. In particular, the context for each term t in a tweet collection $\mathcal{T}$ is represented as a vector $\vec{c} = (c_1, c_2, ..., c_n)$ of terms that occur with t in any tweet in $\mathcal{T}$. The contextual sentiment of t is then extracted by first transforming the term vector c into a 2d circle representation, and then extracting the geometric median of the points (context terms) within the circle. The position of the median within the circle represents the overall contextual sentiment of t. This simple technique has proven effective in calculating contextual sentiment [22].

Figure 2 depicts the representation and extraction of the contextual sentiment of the term "great" by the SentiCircle approach. First, given a tweet collection $\mathcal{T}$, the target term m_{great} is represented as a vector $\vec{c}_{great} = (c_1, c_2, ..., c_n)$ of terms co-occurring with term m in any tweet in $\mathcal{T}$ (e.g., "pain", "loss", ..., "death"). Secondly, the context vector $\vec{c}_{great}$ is transformed into 2d circle representation. The center of the circle represents the target term m_{great} and points within the circle denote the context terms of m_{great}. The position (x_{c_i}, y_{c_i}) of each context term $c_i \in \vec{c}_{great}$ is defined as:

$$x_{c_i} = r_i \cos\theta_i \qquad y_{c_i} = r_i \sin\theta_i \tag{1}$$

Where the angle θ_i represents the prior sentiment of the context term c_i multiplied by π, and it is obtained from the lexicon to be adapted. The radius r_i represents co-occurrence frequency between c_i and the target term m_{great} and it is computed based on the *TF-IDF* weighting scheme as follows:

$$r_i = corr(m_{great}, c_i) = f(c_i, m_{great}) \times \log \frac{N}{N_{c_i}} \tag{2}$$

where $f(c_i, m_{great})$ is the number of times c_i occurs with m_{great} in tweets, N is the total number of terms, and N_{c_i} is the total number of terms that occur with c_i.

Based on the SentiCircle representation, terms with positive prior sentiment are positioned on the upper half of the circle (e.g., please) and terms with negative prior sentiment are positioned in the lower half (e.g., pain, loss). Term co-concurrence determines the distance (i.e., radius) of these terms with respect to the origin. Thirdly, the geometric median G of the SentiCircle of "Great" is computed (-4 in our example), which constitutes the contextual sentiment of the term.[2]

In the following subsection we describe how to adapt the sentiment lexicon using the contextual sentiment extracted in this step.

The reasons for using SentiCircle for extracting terms' contextual sentiment are threefold. First, unlike other approaches, SentiCircle is built for social media data, and specifically for Twitter data [22]. Secondly, it enables detecting not only the contextual sentiment orientation of words (i.e., *positive*, *negative*, *neutral*), but also the words' contextual sentiment strength (e.g., negative(-3), positive(+4)). This in turn allows for better fine-tuning and adaptation of the sentiment of words in the lexicon. Thirdly, SentiCircle, as explained above, relies on simple, yet effective frequency-based representation of words' context. This representation is easy

[1] Also known as *Statistical Semantics* [34]

[2] We refer the reader to the body of [22] for more details about the SentiCircle approach.

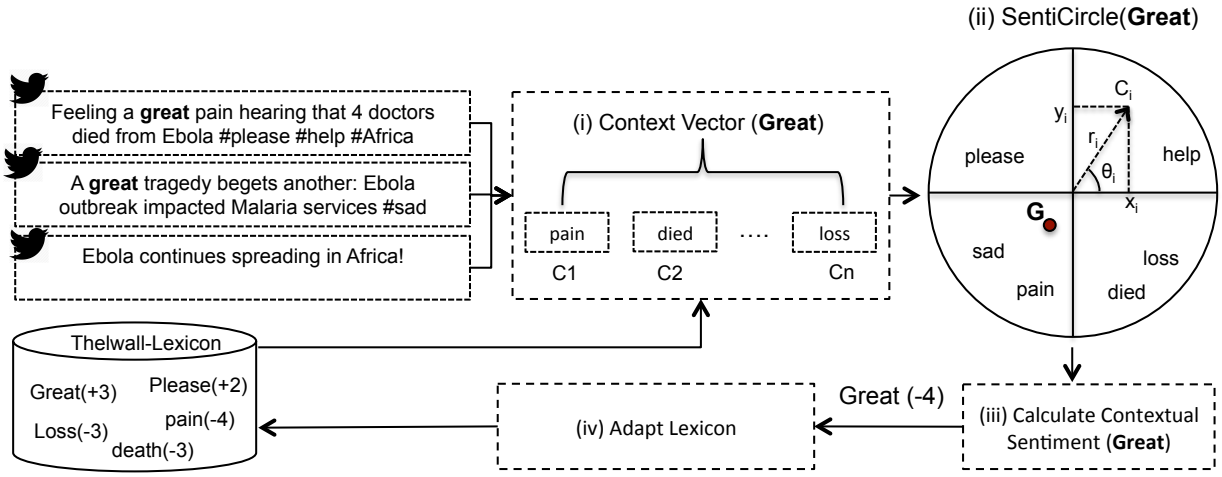


Fig. 2. Illustrative example of extracting the contextual sentiment of the word `Great` from a tweet collection and adapting Thelwall-Lexicon with the new extracted sentiment respectively. Red dot in the SentiCircle represents the geometric median G of the context points in the circle.

to extend and enrich with the conceptual semantics of words, as will be explained in Section 4.

3.2. Rules for Lexicon Adaptation

The second step in our pipeline is to adapt the prior sentiment of terms in a given sentiment lexicon based on the terms' contextual sentiment information extracted in the previous step. To this end, we propose a general rule-based method to decide on the new sentiment of terms in the lexicon. In the following we give a formal definition of the general purpose sentiment lexicon and its properties, and explain how our proposed method functions on it accordingly.

General-purpose sentiment lexicon: is a set of terms $\mathcal{L} = \{t1, t2, ..., t_n\}$ of fixed size n. Each term $t \in \mathcal{L}$ is coupled with a prior sentiment score that is often a numerical value $prior_m \in [-\lambda, -\delta, \delta, \lambda]$, denoting the sentiment orientation and strength of t. In particular, t is *positive* if $prior_t \in (\delta, \lambda]$, *negative* if $prior_t \in [-\lambda, -\delta)$, and *neutral* if $prior_t \in [-\delta, \delta]$. $|\lambda|$ is the maximum sentiment strength that a term can have. The closer the $prior_t$ is to λ the higher the sentiment strength is. $|\delta|$ defines the boundaries of the neutral sentiment range. The values of both, λ and δ depend on the specifications of the studied sentiment lexicon and are defined at the design/construction phase of the lexicon (see section 5.1).

Lexicon adaptation rules: our proposed method uses a set of 4 antecedent-consequent rules (Table 1) to decide how to update the prior sentiment of a term ($prior_t$) in a given lexicon with respect to its contextual sentiment ($contextual_t$). As noted in Table 1, these rules are divided into:

- **Updating rules**: for updating only the existing terms in the lexicon. These rules are further divided into rules that deal with terms with the same prior and contextual sentiment orientations (e.g., both, $prior_t$ and $contextual_t$ are positive or negative), and rules that deal with terms that have different prior and contextual sentiment orientations (i.e., $prior_t$ is negative and $contextual_t$ is positive or vice-versa).
- **Expanding rules**: rules for expanding the lexicon with new opinionated terms.

The notion behind the proposed rules is rather simple: *For a given term $t \in \mathcal{L}$, check how strong/weak the contextual sentiment ($contextual_t$) is and how strong/weak the prior sentiment ($prior_t$) is $\rightarrow$ update $prior_t$ in the lexicon accordingly*. As mentioned earlier, $contextual_t$ is obtained as described in Section 3.1 and its value range $[-\lambda, \lambda]$. The threshold θ is computed as $\theta = |\lambda|/2$ and it is used to determine how *strong/weak* the sentiment of the term is. If the term does not exist in the lexicon, we add it to the lexicon with its corresponding contextual sentiment.

In Thelwall-Lexicon [29], as will be explained in Section 5.1, $|\lambda| = 5$ and $|\delta| = 1$, i.e., the prior sentiment for the terms in this lexicon is between $[-5, +5]$,

Id	Antecedents	Consequent
Updating Rules (Same Sentiment Orientations)		
1	$(\|contextual_t\| > \|prior_t\|) \wedge (\|contextual_t\| > \theta)$	$prior_t = \begin{bmatrix} prior_t + \alpha; prior_t > 0 \\ prior_t - \alpha; prior_t < 0 \end{bmatrix}$
Updating Rules (Different Sentiment Orientations)		
2	$(\|contextual_t\| > \theta) \wedge (\|prior_t\| \leqslant \theta)$	$prior_t = \begin{bmatrix} \alpha; prior_t < 0 \\ -\alpha; prior_t > 0 \end{bmatrix}$
3	$(\|contextual_t\| > \theta) \wedge (\|prior_t\| > \theta)$	$prior_t = \begin{bmatrix} prior_t - \alpha; prior_t > 0 \\ prior_t + \alpha; prior_t < 0 \end{bmatrix}$
Expanding Rule		
4	$term(t) \notin lexicon(\mathcal{L})$	$(prior_t = contextual_t) \wedge AddTerm(t, \mathcal{L})$

Table 1

Adaptation rules for sentiment lexicons, where $AddTerm(t, mathcal L)$: add term t to lexicon $\mathcal{L}$.

and the neutral sentiment range is between $[-1, 1]$. The value of θ is set up to 3.[3]

We use the same example depicted in Figure 2 to show how these rules are applied for lexicon adaptation. The word "Great" in Thelwall-Lexicon has a weak positive sentiment ($prior_{great} = +3; |prior_{great}| = 3 \leqslant \theta$), while its contextual sentiment, as previously explained, is strongly negative ($contextual_{great} = -4; |contextual_{great}| = 4 > \theta$). Therefore, rule number 2 is applied, since the prior and contextual sentiment have different sentiment orientation (i.e., the prior is positive and the contextual is negative). The new prior for the word "Great" will therefore be set up to -$alpha$. In the case of Thelwall-lexicon $alpha$ is 1 (i.e., $adaptedprior_{great} = -1$). In the same example in Figure 2, the word "tragedy" is not covered by the Thelwall-Lexicon, and therefore, it has no prior sentiment. However, its contextual sentiment, extracted using the process described in the previous section, is negative (i.e., $contextual_{tragedy} = -3$). In this case rule number 4 is applied and the term is added to the lexicon with a negative sentiment strength of -3.

4. Semantic Enrichment for Context-based Lexicon Adaptation

In this section we propose enriching our original context-based adaptation model, described in the previous section, with the conceptual semantics of words in tweets. To this end, we follow two different methodologies: (1) Enriching the adaptation model with the semantic concepts of named-entities extracted from a given tweet collection, *Conceptually-enriched Model*. (2) Adjusting the contextual correlation between two co-occurring named-entities in tweets based on the semantic relations between them, *Semantically-adjusted Relations Model*. In the following subsections we describe both enrichment models and the motivation behind them.

4.1. Conceptually-enriched Adaptation Model

In Section 3 we showed our proposed method to adapt sentiment lexicons based on the contextual sentiment of terms in a given collection of tweets. However, relying on the context only for detecting terms' sentiment might be insufficient. This is because the sentiment of a term may be conveyed via its conceptual semantics rather than by its context [5]. In the example in Figure 2, the context of the word "Ebola" in "Ebola continues spreading in Africa!" does not indicate a clear sentiment for the word. However, "Ebola" is associated with the semantic type (concept) "Virus/Disease", which suggests that the sentiment of "Ebola" is likely to be negative.

In order to address to above issue, we propose enriching the context-based lexicon adaptation model with the conceptual semantics of words in tweets. To this end, we add two additional steps to our original pipeline (Figure 1:b): *conceptual semantic extraction*, and *conceptual semantic enrichment*. These two steps are executed prior to the extraction of words' contextual sentiment, as follows:

1. **Conceptual semantic extraction**: This step extracts the named entities that appear in a tweet collection (e.g., "Obama", "Illinois", "NBC") along with their associated semantic types ("Person", "City", "Company") and their

[3] Since Thelwall-Lexicon uses discrete and not continuous values for priors, θ is rounded up to the nearest integer value to match the annotation format of Thelwall-Lexicon

semantic subtypes (e.g., "Politician", "US County", "TV Network"). To this end, we use the semantic extraction tool AlchemyAPI[4] due to its accuracy and high coverage of semantic types and subtypes [23].

2. **Conceptual semantic enrichment**: This step incorporates the conceptual semantics extracted from the previous step into the extraction process of the terms' contextual sentiment. To this end, the entities' semantic subtypes are first added as additional unigrams to the tweets in which the entities occur. After that, the enriched Twitter dataset is passed to the contextual sentiment extraction step, as depicted in Figure 1:b. As mentioned in Section 3.1, the context of a term t in the latter step, is represented as a vector $\vec{c} = (c_1, c_2, ..., c_n)$ of terms that occur with t in a given tweet collection. Using the semantically enriched Twitter dataset to construct the context vector $\vec{c}$ results in extending $\vec{c}$ with the semantic subtypes $\vec{s} = (s_1, s_2, ..., s_m)$ of named entities $\vec{e} = (e_1, e_2, ..., e_m)$ that occur with t in the tweet collection as:

$$\vec{c_s} = \vec{c} + \vec{s} = (c_1, c_2, ..., c_n, s_1, s_2, ..., s_m) \quad (3)$$

where $\vec{c_s}$ is the new semantically-enriched contextual vector of t, which will be subsequently used instead of $\vec{c}$ to extract the overall contextual sentiment of t.

Note that we currently rely only on the entities' semantic subtypes for the semantic enrichment phase, excluding the semantic types. Unlike semantic types, semantic subtypes capture more fine-grained knowledge about the entity (e.g., "Obama" > "Politician").

4.2. Semantically-adjusted Relations Model

Using the distributional semantic hypothesis, our context-based approach assigns a stronger relation to words that tend to co-occur more frequently in same context. However the document collection may represent only a partial view of the contexts in which two words my co-occur together. For example, in the GASP Twitter dataset around the dialogue for earth gas prices[1], the entities *Barack Obama* and *Texas* tend to appear together and therefore have a strong contextual relation. However, these two entities are related within a high number of different contexts. Figure 3 shows a small sample of the different semantic contexts that link the two previous entities. These contexts include Barack Obama's birth place, his candidatures and his duties as president.

To capture the variety of contexts in which two terms can potentially appear together we compute the number of relations between these two terms in DBPedia by using the approach proposed by Pirro [19]. Our assumption is that the strength of the contextual relation between two terms, captured by their co-occurrence within the document collection, should be modified according to the number of contexts in which these terms can potentially appear together. The smaller the number of contexts, the stronger the contextual relation should be.

Based on the above assumption we propose adjusting the strength of the contextual relations between terms, captured by the context-based model, by using the semantic relations between them. To this end, we add two additional steps to the original pipeline (see Figure 1:c): *semantic relation extraction* and *semantic relation adjustment*. These two steps are further described below.

1. **Semantic relation extraction**: This step extracts the sets of semantic relations for every pair of named entities co-occurring together in the tweets. For the purpose of our study we extract semantic relations using the approach proposed by Pirro [19] over DBPedia, since DBPedia is a large generic knowledge graph which captures a high variety of relations between terms. To extract the set of relations between two name entities this approach takes as input the identifiers (i.e., URIs) of the source entity e_s, the target entity e_t and an integer value K that determines the maximum path length of the relations between the two named entities. The output is a set of SPARQL queries that enable the retrieval of paths of length at most K connecting e_s and e_t. Note that in order to extract all the paths, all the combinations of ingoing/outgoing edges must be considered. Following our previous example, if we were interested in finding paths of length $K <= 2$ connecting $e_s = Obama$ and $e_t = Texas$ our approach will consider the following set of SPARQL queries:
SELECT * WHERE {:Obama ?p1 :Texas}
SELECT * WHERE {:Texas ?p1 :Obama}
SELECT * WHERE {:Obama ?p1 ?n1. ?n1 ?p2 :Texas}
SELECT * WHERE {:Obama ?p1 ?n1. :Texas ?p2 ?n1}
SELECT * WHERE {?n1 ?p1 :Obama. :Texas ?p2 ?n1}
SELECT * WHERE {?n1 ?p1 :Obama. ?n1 ?p2 :Texas}

[4] www.alchemyapi.com

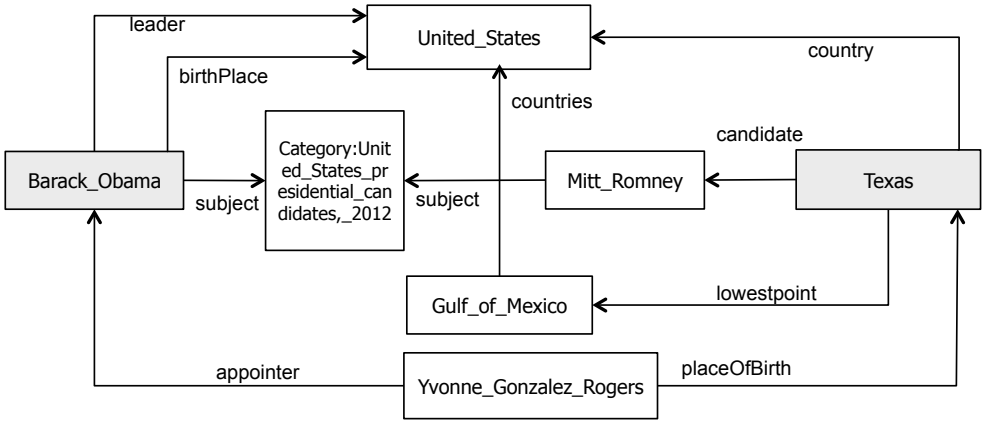


Fig. 3. Example for sentiment relations between the entities *Barack Obama* and *Texas* with a path length of ≤ 3

As it can be observed, the first two queries consider paths of length one. Since a path may exist in two directions, two queries are required. The retrieval of paths of length 2 requires 4 queries. In general, given a value K, to retrieve paths of length K, 2^k queries are required.

2. **Semantic relation adjustment**: Now we have for every pair of named entities (e_s, e_t) co-occurring together in the tweets, a set $\mathcal{R}_{(e_s,e_t)} = \{p1, p2, ..., p_N\}$ of paths of size N, representing the semantic relations between e_s and e_t.
As mentioned earlier, our goal behind enriching the context-based model with semantic relations is to adjust the strength of the contextual relation between e_s and e_t based the number of semantic relations (paths) between them. To this end, we construct the SentiCircle S_{e_s} of the source entity e_s, as depicted in Figure 4. Since both entities co-occur together in tweets, the target entity e_t is positioned in the SentiCircle S_{e_s} with a radius r_t representing the strength of the contextual relation between e_s and e_t, as described in Section 3.1. Therefore, the task of adjusting the contextual relations between e_s and e_t breaks down into altering the value of r_t as follows:

$$r'_t = r_t + \left[\frac{N}{M}(1 - r_t)\right] \quad (4)$$

Where N is the number of the semantic paths between e_s and e_t extracted in the previous step, M is the maximum number of paths extracted for a pair of entities in the Twitter dataset, and r'_t is the new radius of entity e_t after adjustment.

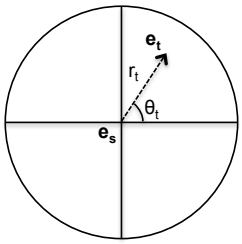


Fig. 4. SentiCircle of entity e_s showing the strength of the contextual relation between e_s and e_t, represented by the radius r_t

As can be noted, the above equation modifies the value of r_t based on the number of paths between e_s and e_t. The smaller the number of paths is, the stronger the contextual relation should be, and thereby the higher the value of r'_t is.
Note that the enrichment by semantic relations in this model is done in two iterations of the adaptation process. Specifically, in the first iteration the sentiment lexicon is adapted using the original context-based model (Figure 1:a). In the second iteration the semantically-adjusted relation model is applied on the adapted lexicon, where the semantic-relation adjustment takes place. Adaptation in the first iteration allows us to capture the contextual relations of entities within tweets and assign them a sentiment value. Note that sentiment lexicons are generic and most of the tweet entities (e.g., `Obama`, `Texas`) will not appear in these lexicons. By relying on one iteration of adaptation only, an entity will have little impact on the contextual sentiment of other entities since entities don't generally have any initial sentiment score within the lexicon to be adapted. Hence, a second iteration of adaptation is required in order

to detect the sentiment of entities that do not occur in the lexicon, and maximise the impact of the semantic relation adjustment in our models.

5. Experimental Setup

In this section we present the experimental set up used to assess our proposed lexicon adaptation models, the context-based adaptation model (Section 3) and the semantic adaptation models (Section 4). This setup requires the selection of: (i) the sentiment lexicon to be adapted, (ii) the context (Twitter datasets) for which the lexicon will be adapted, (iii) the baseline models for cross-comparison, (iv) the different configurations for adapting the lexicon and, (iv) the semantic information used for the semantic adaptation models. All these elements will be explained in the following subsections. We evaluate the effectiveness of our method by using the adapted lexicons to perform tweet-level sentiment detection, i.e., detect the overall sentiment polarity (positive, negative) of tweets messages.

5.1. Sentiment Lexicon

For the evaluation we choose to adapt the state-of-the-art sentiment lexicon for social media; Thelwall-Lexicon [30,29]. Thelwall-Lexicon is a general purpose sentiment lexicon specifically designed to function on social media data. It consists of 2546 terms coupled with values between -5 (very negative) and +5 (very positive), defining their sentiment orientation and strength. Terms in the lexicon are grouped into three subsets of 1919 negative terms ($prior_t \in$[-2,-5]), 398 positive terms ($prior_t \in$[2,5]) and 229 neutral terms ($prior_t \in$[-1,1]). Based on the aforementioned specifications, the parameters in our proposed adaptation method (Section 3.2) are set as: $|\lambda| = 5$, $|\delta| = 1$, $|\theta| = 3$ and $|\alpha| = 1$. Thelwall-lexicon was selected for this evaluation because, to the best of our knowledge, it is currently one of the best performing lexicons for computing sentiment in social media data.

5.2. Evaluation Datasets

To assess the performance of our lexicon adaptation method we require the use of datasets annotated with sentiment labels. for this work we selected three evaluation datasets often used in the literature of sentiment analysis (SemEval, WAB and GASP) [21]. These datasets differ in their sizes and topical focus. Numbers of positive and negative tweets within these datasets are summarised in Table 2.

5.3. Evaluation Baseline

As discussed in Section 2, several methods have been proposed for context-based sentiment lexicon bootstrapping and/or adaptation. In this paper we compare our adaptation models against the semantic orientation by association approach (*SO*) [32], due to its effectiveness and simple implementation. To generate a sentiment lexicon, this approach starts with a balanced set of 14 positive and negative paradigm words (e.g., good, nice, nasty, poor). After that, it bootstraps this set by adding words in a given corpus that are statistically correlated with any of the seed words. The new words added to the lexicon have positive orientation if they have a stronger degree of association to positive words in the initial set than to negative ones, and vice-versa. Statistical correlation between words is measured using the *pointwise mutual information* (PMI). From now on we refer to this approach shortly as the *SO-PMI method*.

We apply the SO-PMI method to adapt Thelwall-Lexicon to each of the three datasets in our study in the same manner as described above. Specifically, given a twitter dataset we compute the pointwise mutual information between each opinionated term t in Thelwall-Lexicon and each word w that co-occur with t in the the twitter dataset as follows:

$$PMI(t,w) = \log\left(\frac{p(t,w)}{p(t)p(w)}\right) \tag{5}$$

After that, we assign w the sentiment orientation (SO) of the term in Thelwall-Lexicon ($\mathcal{L}$) that have the highest PMI with w as:

$$SO(w) = SO\left(\arg\max_{t\in\mathcal{L}} PMI(t,w)\right) \tag{6}$$

Note that in addition to the lexicons adapted by the SO-PMI method, we also compare our proposed approaches against the original Thelwall-Lexicon without any adaptation.

5.4. Configurations of the Lexicon Adaptation Models

We test our context-based adaptation model and the two semantic adaptations model under three different configurations. We use terms from the running example in Figure 2 to illustrate the impact of each adaptation model:

1. *Lexicon Update (LU)*: The lexicon is adapted only by updating the prior sentiment of existing

Dataset	Tweets	#Negative	#Positive	#Unigrams
Semeval Dataset (SemEval) [17]	7520	2178	5342	22340
The Dialogue Earth Weather Dataset (WAB) [1]	5482	2577	2905	10917
The Dialogue Earth Gas Prices Dataset (GASP) [1]	6032	4999	1033	12868

Table 2
Twitter datasets used for evaluation. Details on how these datasets were constructed and annotated are provided in [21].

	SemEval	WAB	GASP
No. of Entities	2824	685	750
No. of Semantic Types (Concepts)	31	25	23
No. of Semantic Subtypes	230	93	109

Table 3
Unique Entity/Types/Subtypes for the SemEval, WAB and GASP datasets

Dataset	numRelations	minPath	maxPath	AveragePath
SemEval	1,011,422	1	3	2.82
GASP	811,741	1	3	2.95
WAB	796,021	1	3	2.92

Table 5
Amount of relations and path lengths extracted for each dataset

terms. In our running example, the prior sentiment of the pre-existing word "`Great`" in Thelwall-Lexicon (i.e., $prior_{great}$=+3) will be updated based on the words' contextual sentiment (i.e., $contextual_{great} = -4$) to -1.

2. *Lexicon Expand (LE)*: The lexicon is adapted only by adding new opinionated terms. Here new words, such as "`Tragedy`" and "`Ebola`", along with their contextual sentiment, will be added to the lexicon.
3. *Lexicon Update and Expand (LUE)*: The lexicon is adapted by adding new opinionated terms ("`Tragedy`" and "`Ebola`") and by updating the prior sentiment of existing terms ("`Great`").

5.5. Extracted Semantics

We use AlchemyAPI to extract the conceptual semantics of named entities from the three evaluation datasets (Section 4.1). Table 3 lists the total number of entities extracted and the number of semantic types and subtypes mapped against them for each dataset. Table 4 shows the top 10 frequent semantic subtypes under each dataset. As mentioned in Section 4.1, we only use the entities' semantic subtypes for our semantic enrichment, mainly due to their stronger representation and distinguishing power than general higher level types (e.g., "Person"). Table 5 shows the number of semantic relations extracted between the entities of each dataset. This table also includes the minimum, maximum and average path length among all the extracted relations. A maximum path length of 3 was consider for our experiments.

6. Evaluation Results

In this section, we report the results obtained from using the different adaptations of Thelwall-Lexicon to compute tweet-level sentiment detection. To compute sentiment, we use the approach proposed by Thelwall [29], where a tweet is considered positive if its aggregated positive sentiment strength (i.e., the sentiment strength obtained by considering the sentiment weights of all words in the tweet) is 1.5 times higher than the aggregated negative one, and vice versa. Our baselines for comparison are the original version of Thelwall-Lexicon and the version adapted by the SO-PMI method. Results in all experiments are computed using 10-fold cross validation over 30 runs of different random splits of the data to test their significance. Statistical significance is done using *Wilcoxon signed-rank test* [28]. Note that all the results in F1-measure reported in this section are statistically significant with $\rho < 0.001$. Evaluation in the subsequent sections consists of 5 main phases:

1. Measure the performance of our context-based adaptation model using the three evaluation datasets and the three adaptation settings (Section 6.1).
2. Evaluate the performance of the conceptually-enriched adaptation model and report the evaluation results averaged across the three datasets (Section 6.2).
3. Test the performance of the semantically-adjusted relations model on the three evaluation datasets (Section 6.3).
4. Conduct a statistical analysis on the impact of our adaptation models on Thelwall-Lexicon (Section 6.5).

SemEval		WAB		GASP	
Subtype	Frequency	Subtype	Frequency	Subtype	Frequency
TVActor	505	AdministrativeDivision	93	AwardWinner	350
AwardWinner	351	GovernmentalJurisdiction	91	Politician	328
MusicalArtist	344	Location	66	Celebrity	321
Filmactor	324	Placewithneighborhoods	49	Location	104
Athlete	316	PoliticalDistrict	45	AdministrativeDivision	103
Location	263	Sportsteam	12	GovernmentalJurisdiction	102
GovernmentalJurisdiction	263	FieldofStudy	11	PlaceWithNeighborhoods	15
Footballplayer	238	Invention	10	Musicalartist	14
Celebrity	230	MusicalArtist	10	AutomobileCompany	13
AwardNominee	225	VentureFundedCompany	9	BroadcastArtist	11

Table 4

Top 10 frequent semantic subtypes of entities extracted from the three datasets

5. Study the effect of the sentiment class distribution on the performance of our adaptation models (Section 6.6).

6.1. Results of Context-based Lexicon Adaptation

The first task in our evaluation is to assess the effectiveness of our context-based adaptation model. Table 6 shows the results of binary sentiment detection of tweets performed on the three evaluation datasets using (i) the original Thelwall-Lexicon (*Original*), (ii) Thelwall-Lexicon adapted under the update setting (LU), (iii) Thelwall-Lexicon adapted under the expand setting (LE), and (iv) Thelwall-Lexicon adapted under the update and expand setting (LUE). The table reports accuracy and three sets of precision (P), recall (R), and F1-measure (F1), one for positive sentiment identification, one for negative sentiment identification, and the third shows the averages of the two.

In the case of the SemEval and GASP datasets the LU and LUE lexicons outperform the original lexicon in all the average measures by up to 6.5%. For example, LU and LUE on SemEval improve the performance upon the original lexicon by at least 6.3% in accuracy and 5.1% in average F1. Similarity, the improvement of LU and LUE on GASP reaches 5% and 4.6% in accuracy and F1 comparing to the original lexicon. Adapting lexicons by expanding terms only (LE) does not have much impact on the sentiment detection performance.

Compared to the state-of-the-art SO-PMI method, we notice a similar performance trend for all the lexicons. In particular, the SO-PMI lexicon outperforms the original lexicon on both, the SemEval and GASP datasets by up to 2.8% in accuracy and 2.5% in average F1. However, both the LU and LUE lexicons outrun the the SO-PMI lexicon by 3.9% in accuracy and 3.3% in average F1. The LE lexicon, on the other hand, gives on average 1.7% and 1.5% lower performance in Accuracy and F1 than the SO-PMI, respectively.

In the case of the WAB dataset, the highest sentiment detection performance (79.24% in accuracy and 79.16% in average F1) is obtained using the original lexicon. In this case, the context-based adaptation model has a modest impact. Only the LU lexicon on WAB gives a 0.2% better precision than the original lexicon. Compared to SO-PMI, our adaptation model gives a comparable performance, except for the case of the LE lexicon, where the performance improves by 0.3% in average F1.

Overall, the average performance across the three datasets shows that the improvement of the adapted LU and LUE lexicons over the original lexicon and the SO-PMI lexicon reaches 3.9% and 2.9% in accuracy, and 3.2% and 2.4% in F1 respectively. On the other hand, the LE lexicon gives negligible performance improvements over the original lexicon, and a slightly lower performance than SO-PMI by 0.9% and in accuracy and 0.6% in F1.

The variation in the performance of our adapted lexicons throughout the three datasets might be due to their different sentiment class distribution. According to Table 2 the class distribution in SemEval and GASP is highly skewed towards the positive and negative classes respectively. On the other hand, the WAB dataset is the most balanced dataset amongst the three. The impact of such skewness on sentiment detection is investigated further in Section 6.6.

6.2. Results of the Conceptually-enriched Adaptation Model

The second evaluation task in this paper is to assess the effectiveness of our conceptually-enriched model in adapting sentiment lexicons (Section 4.1). Table 7 shows the *average results across the three datasets* considering the three different settings of lexicon adaptation: *update*, *expand*, and *update and expand*. We re-

Dataset	Lexicon	Accuracy	Negative Sentiment			Positive Sentiment			Average		
			P	R	F1	P	R	F1	P	R	F1
SemEval	Original	71.85	50.84	85.17	63.67	91.66	66.42	77.02	71.25	75.79	70.35
	SO-PMI	73.9	53.19	82.37	64.64	90.74	70.44	79.31	71.96	76.41	71.97
	LU	**76.9**	57.35	79.02	66.46	89.89	76.04	82.39	**73.62**	**77.53**	**74.42**
	LE	72.14*	51.16	83.79	63.53	91.07	67.39	77.46	71.12	75.59	70.5
	LUE	76.66	57.07	78.42	66.06	89.62	75.95	82.22	73.34	77.18	74.14
WAB	Original	**79.24**	77.96	77.84	77.9	80.37	80.48	80.43	79.17	79.16	**79.16**
	SO-PMI	79.04	80.49	73.15	76.64	77.96	84.27	80.99	79.22	78.71	78.82
	LU	79.11	81.05	72.53	76.55	77.71	84.96	81.17	**79.38**	78.74	78.86
	LE	79.15	*78.03*	77.45	77.74	80.13	80.65	*80.39*	79.08	79.05	79.07
	LUE	79	80.87	72.49	76.45	77.65	84.78	81.06	79.26	78.64	78.75
GASP	Original	69.38	86.89	74.25	80.08	26.88	45.79	33.87	56.88	60.02	56.97
	SO-PMI	69.69	86.73	74.89	80.38	26.82	44.53	33.48	56.77	59.71	56.93
	LU	**73.01**	87.46	78.72	82.86	30.59	45.4	36.55	**59.03**	**62.06**	**59.71**
	LE	69.21	86.9	74.01	79.94	26.78	45.98	33.84	56.84	60	56.89
	LUE	72.99	87.44	78.72	82.85	30.55	45.3	36.49	59	62.01	59.67
Average	Original	73.49	71.90	79.09	73.88	66.30	64.23	63.77	69.10	71.66	68.83
	S0-PMI	74.21	73.47	76.80	73.89	65.17	66.41	64.59	69.32	71.61	69.24
	LU	**76.34**	75.29	76.76	75.29	66.06	68.80	66.70	**70.68**	**72.78**	**71.00**
	LE	73.50	72.03	78.42	73.74	65.99	64.67	63.90	69.01	71.55	68.82
	LUE	76.22	75.13	76.54	75.12	65.94	68.68	66.59	70.53	72.61	70.85

Table 6

Results obtained from adapting Thelwall-Lexicon on three datasets using the context-based adaptation model. Bold=highest performance. Italic=significance at 0.05, None-Italic=significance < 0.001, *=significance at 0.1

fer to adapted lexicons by the conceptually-enriched model under these settings as SLU, SLE, and SLUE to differentiate them from the lexicons adapted by the context-based model. Note that here we do not discuss the results of the semantic model on each dataset to avoid repetition, as the performance trend of the semantic model on each of the datasets is very similar to the one reported for the context-based model. For the complete list of results we refer the reader to Table 11 in Appendix A.

As we can see in the table, the original lexicon gives the lowest performance in all measures in comparison with the SO-PMI lexicon and the conceptually adapted lexicons, SLU, SLE and SLUE. In particular, the SLU lexicon achieves the highest performance among all other lexicons, outperforming the original lexicon by 4.1% in accuracy and 3.3% in average F1. The SLUE lexicon comes next with quite close performance to the SLU lexicon; SLUE produces 4.0% and 3.2% higher accuracy and F1 than the original lexicon respectively. The SLE lexicon comes third with marginal impact on sentiment detection performance.

Compared to the SO-PMI lexicon, a similar performance trend of our conceptually-adapted lexicons can be observed. Specifically, both, SLU and SLE outperform SO-PMI by up to 2.96% in accuracy and 2.6% in average F1. On the other, the SLE lexicon produces lower performance than the SO-PMI lexicon by 0.56% and 0.43% in accuracy and average F1 respectively.

6.3. Results of the Semantically-adjusted Relations Model

The third step in our evaluation is to test the performance of the adapted lexicons by the semantically-adjusted relations model (Section 4.2). The lower part of Table 7 lists the average results across the three datasets for the adapted lexicon under the update setting (SRU), the expand setting (SRE), and the update and expand setting (SRUE). For the complete list of results we refer the reader to Table 12 in Appendix A.

According to these results in Table 7, we notice that the three semantically adapted lexicons SRU, SRE and SRUE outperforms both, the original lexicon and the SO-PMI lexicon by a large margin. In particular, the lexicon adapted under the expand setting, SRE outperforms both the baseline lexicons by up to 4.3% in accuracy and 3.2% in average F1. The SRU and the SRUE lexicons come next by a performance that is 3.6% and 2.2% higher in accuracy and F1 than the baselines respectively.

Model	Lexicon	Accuracy	Negative Sentiment			Positive Sentiment			Average		
			P	R	F1	P	R	F1	P	R	F1
Baselines	Original	73.49	71.90	79.09	73.88	66.30	64.23	63.77	69.10	71.66	68.83
	PMI	74.21	73.47	76.80	73.89	65.17	66.41	64.59	69.32	71.61	69.24
Conceptually-enriched Model	SLU	**76.47**	75.54	76.31	75.29	65.93	69.20	66.87	**70.74**	**72.75**	**71.08**
	SLE	73.78	72.40	77.02	73.58	65.44	65.66	*64.31*	68.92	71.34	68.94
	SLUE	76.42	75.47	76.31	75.24	65.90	*69.10*	66.81	*70.69*	72.71	71.03
Semantically-adjusted Relations Model	SRU	76.14	75.99	74.74	74.50	64.78	68.95	66.14	70.39	71.84	70.31
	SRE	**76.66**	77.38	73.62	74.76	64.84	71.42	67.31	**71.11**	**72.52**	**71.03**
	SRUE	76.13	76.03	74.75	74.52	64.78	69.01	66.16	70.41	71.88	70.34

Table 7

Average results across the three datasets of Thelwall-Lexicon adapted by the semantic model. Italic=significance at 0.05, None-Italic=significance < 0.001

6.4. Context-based Adaptation vs. Semantic-based Adaptation

In the previous sections we showed that lexicons adapted by our context-based model as well as both the semantically-enriched models outperform both the original lexicon and the SO-PMI lexicon in most evaluation scenarios.

In this section we investigate how the conceptually-enriched model and the semantically-adjusted relations model perform in comparison with the original context-based adaptation model. Such comparison allows us to understand and highlight the added value of using word semantics for sentiment lexicon adaptation. To this end, we compute the win/loss in accuracy, P, R and average F1 when using both semantic models for lexicon adaptation compared to the context-based model across the three datasets, as depicted in Figure 5.

The results show that the impact of the two semantic models varies across the three lexicon adaptation settings. Specifically, we notice that under both, the *lexicon update setting* and the *lexicon update & expand setting* (Figures 5:a and 5:c) the conceptually-enriched model improves performance upon the context-based model in accuracy, P, and F1 by up to 0.27%, but only gives similar recall. On the other hand, the semantically-adjusted relations model always gives, under these settings, a lower performance on all measures compared to the context-based model.

A different performance trend can be noted for the *expand setting* (Figure 5:b). While conceptually-enriched model has no significant improvement over the context-based model, the semantically-adjusted relation model boosts the performance substantially, with 4.12% and 3.12% gain in accuracy and F1 respectively.

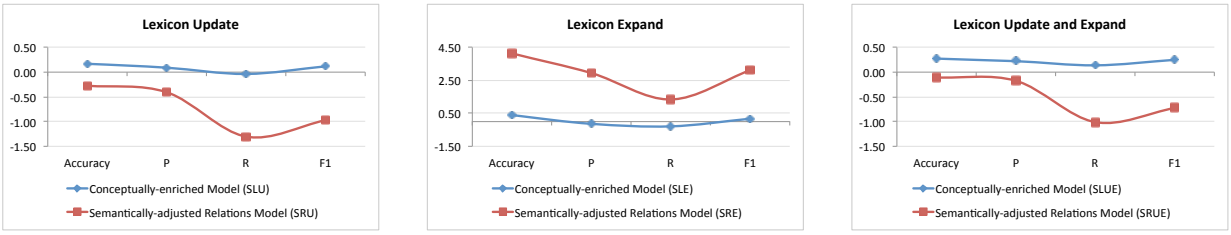
Hence, we can notice that while both semantic enrichment models have a noticeable impact on the lexicon adaptation performance, the conceptually-enriched model has a higher impact on tuning the sentiment of existing words in the lexicon (i.e., *the update setting*). On the other hand, the semantically-adjusted relations model is more useful in expanding the lexicon with new opinionated words (i.e., *the expand setting*). This is probably due to the mechanism in which each model functions. As described in Section 4, the enrichment with semantic concepts is done at the dataset level (Figure 1:b) in the first iteration of the lexicon adaptation process. Contrarily, the enrichment with semantic relations is performed during the contextual-relation extraction phase in the second iteration of the lexicon adaptation process. This will be further discussed in Section 7.

6.5. Adaptation Impact on Thelwall-Lexicon

Applying our adaptation models to Thelwall-Lexicon results in substantial changes to the lexicon. Table 8 shows the average percentage of words across the three datasets that, either changed their sentiment polarity and strength, or were added to the lexicon, by both, context-based adaptation model and the conceptually-enriched model.[5]

On average only 5% of the words in the datasets were found in the original lexicon. However, adapting the lexicon by either model resulted in 38% of these words flipping their sentiment orientation and 60% changing their sentiment strength while keeping their prior sentiment orientation. Only 1% of the words that were found in Thelwall-Lexicon remained untouched. Also, 10% and 4% of previously unseen (hidden) words in the original lexicon were assigned positive and negative sentiment, and were added to the adapted lexicons accordingly. Adding semantic information helped detecting more words in the original lexicon as well as adding more positive and negative terms to the adapted

[5]Note that we do not report statistics for the semantically-adjusted relations model in Table 8, since they are similar to the once of the conceptually-enriched model.



(a) Lexicon Update (b) Lexicon Expand (b) Lexicon Update and Expand

Fig. 5. Win/Loss in Accuracy, P, R and F1 measures of adapting sentiment lexicons by the semantic subtype model and semantic relations model in comparison with the context-based model.

	Context-based Adaptation	Semantic Adaptation	Average
Words found in the lexicon	4.96	5.01	5.00
Words flipped their sentiment orientation	38.40	38.32	38.36
Words changed their sentiment strength	60.60	60.62	60.61
New positive words	10.77	11.03	10.90
New negative words	4.50	4.68	4.59

Table 8

Average percentage of words across the three datasets that had their sentiment orientation or strength updated by the context-based and semantic adaptation models.

Semantic Subtypes	
FilmActor	Positive
GovernmentalBody	Negative
Composer	Positive
Airline	Neutral
Inventor	Positive
Location	Neutral
Physician	Negative
OrganizationSector	Negative
Footballplayer	Neutral
University	Neutral

Table 9

Example of 10 subtypes of entities added to the lexicon after adaptation

lexicon. Table 9 shows an example of 10 semantic subtypes added to the Thelwall-Lexicon by our adaptation model.

6.6. Impact of Sentiment Class Distribution

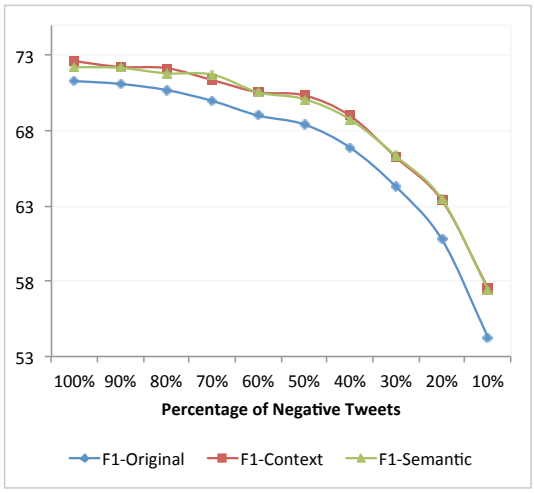
In this section we analyse the impact of sentiment class distribution in the datasets on the performance of our adaptation models. To this end, we first balance the number of positive and negative tweets in the three datasets by mapping the size of the dominant sentiment class to the size of the minor sentiment class, as shown in Table 10. Once we have a balanced

Dataset	**Tweets**	**#Negative**	**#Positive**
SemEval [17]	4356	2178	2178
WAB [1]	5154	2577	2577
GASP [1]	2066	1033	1033

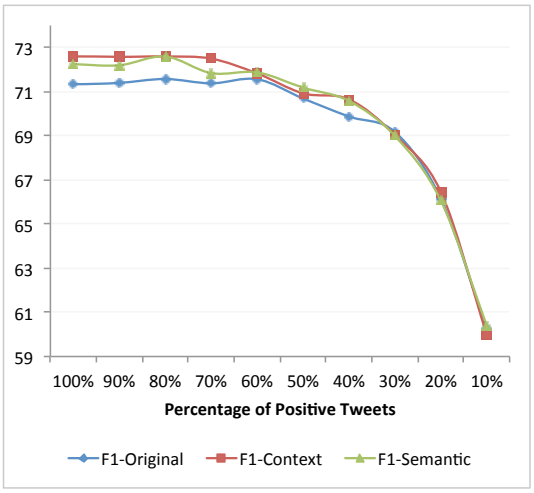
Table 10

Number of positive and negative tweets in the balanced SemEval, WAB and GASP datasets.

dataset with the same number of elements in the positive and negative classes we imbalance this dataset by fixing one class and reducing the number of elements in the other class by 10% in each step (e.g., maintaining all elements of the positive class and reducing the elements of the negative class by 10%, 20%, 30%, ...etc.). By performing this process we obtain 20 versions (folds) of the same dataset, from completely balanced, to completely skewed towards the positive class, to completely skewed towards the negative class. Figure 6 shows the average F1 of binary sentiment detection of applying the original Thelwall-Lexicon (F1-Original), the context-based adapted lexicon (F1-Context) and the semantically adapted lexicon by the conceptually-enriched model (F1-Semantic) on the 20 imbalanced folds of tweets. Note that the results here are averaged over the three datasets and the three adaptation settings.



(a) Positively-skewed Class Distribution



(b) Negatively-skewed Class Distribution

Fig. 6. Average F1 of applying the original and the adapted lexicons on 10 folds of tweets of (a) positively-skewed class distribution and (b) negatively-skewed class distribution.

Figure 6:(a) depicts the performance over the 10 positive-skewed tweet folds. Here we can see that the performance of all lexicons decreases by gradually lowering the number of negative tweets in the data (i.e., increasing the degree positive-skewness). However, we notice that lexicons adapted by both our proposed context-based and semantic adaptation models consistently outperform the original lexicon by 3% in average F1 in all degrees of positive class skewness.

Figure 6:(b) shows the performance over the 10 negatively-skewed folds. We notice that both adapted lexicons keep a 1% higher F1 on average than the original lexicon up to level where number of positive tweets is less than 40% (equals to 60% negative-skewness degree). After that level, all the three lexicons just give similar performance.

It is worth noting that all lexicons, including the original one, are more affected by the positive-skewness in the data than the negative-skewness. Lexicons applied on positively-skewed data give a 2.6% lower F1 on average than lexicons applied on negatively-skewed data. This might be due to imbalanced number of the opinionated words in Thelwall-Lexicon. As mentioned in Section 5.1, Thelwall-Lexicon has 79% more negative words than positive ones.

Overall, one can conclude that the sentiment class distribution clearly impacts the performance of the original lexicon as well as the adapted ones. The more skewed the distribution is (in either direction), the lower the performance is. Nevertheless, results show that lexicons adapted by our models are more tolerant to imbalanced sentiment distributions in the data than the original lexicon. In real life scenarios, imbalanced distributions of tweets' sentiment are perhaps more likely to occur, and lexicon adaptation methods can therefore help enriching sentiment identification in such scenarios.

7. Discussion and Future Work

One of the most fascinating dimensions of social media is the way in which new topics and themes constantly emerge and dissipate. The ability of accurately identifying opinions and sentiment in this dynamic environment is crucial to governments, organisations and business who want to profit from the users' opinions expressed within this medium.

To this end, this paper proposed an approach for adapting general-purpose sentiment lexicons to particular domains or contexts. While our proposed approach is generic, this study focuses on Twitter. However, the use of contextual and semantic information may affect differently the adaptation of sentiment lexicons in different social media platforms (e.g., Facebook, Tumblr), as well as conventional data sources (e.g., online forums, product reviews websites). Further work is therefore needed to study variances across these different types of social and conventional data.

Our selection of the approach to extract the words' contextual sentiment, SentiCircles [22], is inspired by the scope of the provided information, since it does not only capture the words' contextual sentiment orienta-

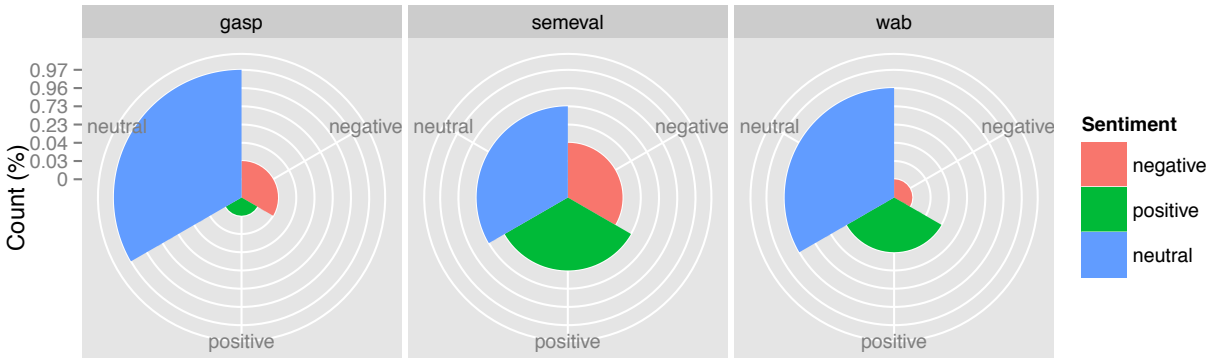


Fig. 7. Percentage of positive, negative and neutral semantic subtypes extracted from the three datasets and added to the lexicon after adaptation

tion but also the words' contextual sentiment strength, which enables a more fine-grained adaptation of the lexicons.

For our experiments we selected to use Thelwall-Lexicon[30] since it is one of the most popular lexicons for sentiment analysis for the social web. However, a more extensive experimentation is needed to assess whether sentiment lexicons of different characteristics may require different types of adaptation.

In our work we have used a third-party commercial tool (AlchemyAPI) to extract the semantic concepts and subtypes of words from tweets. As future work, we plan to experiment with other entity extraction tools, such as DBpedia Spotlight[6] or TexRazor.[7]

Similarly, the approach of Pirro [19] was used to extract semantic relations between every pair of entities within our datasets. Our assumption is that different relations reflect different contexts in which the two words can appear together. However, a better filtering and/or clustering of semantic relations may be needed to provide a more fine-grained identification of these contexts.

In our experiments, we have observed that the use of semantic information helps to improve lexicon adaptation performance (Section 6.4). However, results showed that enriching the adaptation process with semantic subtypes (i.e., the conceptually-enriched adaptation model) did not have much impact on the performance when expanding the lexicon with new opinionated terms. This is probably due to the type of sentiment assigned to the semantic subtypes during the enrichment process. Figure 7 shows the sentiment distribution of the semantic subtypes for the three evaluation datasets. According to this figure, we notice that, on average, 90% of the subtypes added to the lexicon were assigned neutral sentiment after adaptation, while only 9% and 1% of the added subtypes were assigned positive and negative sentiment respectively.

Unlike enrichment with semantic subtypes, the enrichment with semantically-adjusted relations was performed in two iterations of the lexicon adaptation process. The semantic relations between two entities were used to tune the strength of the entities' contextual relations computed in the second iteration (Section 4.2). Such enrichment strategy has proven to enhance the lexicon adaptation performance, especially when expanding lexicons with new opinionated terms (Figure 5:b). As future work, we plan to investigate the case of running the lexicon adaptation process for higher number of iterations, and study the impact of doing so on the lexicon's performance as well as on the run-time complexity of our models.

Extracting semantic relations between a high number of entities via a SPARQL endpoint is a high-cost process. Specific details of the cost of extracting these relations are discussed in [19]. Our implementation uses multithreading, so that queries are sent in parallel to enhance the performance of the retrieval of relations. However, with an increase in maximum path length, the likelyhood of a path existing between two entities increases, as well as the amount of existing paths. In our implementation, we consider a maximum path lenght of 3. Note that higher values of maximum path length come close to the diameter of the DBPedia graph itself and may lead to an explosion in the number of extracted

[6] https://github.com/dbpedia-spotlight/dbpedia-spotlight

[7] https://www.textrazor.com/

relationships[8]. Despite the cost of this step, it is important to consider that this process is computed once per dataset and that relations extracted between entities can be stored and reused when adapting lexicons to Twitter collections of similar topics.

For our evaluation we chose to compare lexicons adapted by our proposed models against the original Thelwall-Lexicon as well as the lexicon adapted by the state-of-the-art SO-PMI method. In our future work we also aim to investigate how our adapted lexicons perform when compared against lexicons generated from scratch by other existing methods. This will provide us with better insights on whether adapting lexicons is preferable, not only in terms of efficiency but also in terms of performance, on on the situations for which one method may be better than the other one.

In summary, while there is still extensive room for future work, our experiments and results show how contextual and semantic information can be combined to successfully adapt generic-purpose sentiment lexicons to specific contexts, helping therefore to correctly identify the sentiment expressed by social media users. We hope that the presented study will serve as bases for future work within the community and enable further research into the semantic adaptation of sentiment lexicons for microblogs.

8. Conclusions

Although much research has been done on creating domain-specific sentiment lexicons, very little attention has been giving to the problem of lexicon adaptation in social media, and to the use of semantic information as a resource to perform such adaptations.

This paper proposed a general method to adapt sentiment lexicons based on contextual information, where the domain or context of adaptation is defined by a collection of posts. A semantic enrichment of this method is also proposed where conceptual semantics are used to better capture the context for which the lexicon is being adapted.

An evaluation of our proposed method was performed by adapting the state-of-the-art sentiment lexicon for the social web [30] to three different contexts (Twitter datasets) using various configurations of our proposed approach. Results showed that the adapted sentiment lexicons outperformed the baseline methods in average by 3.4% in accuracy and 2.8% in F1 measure, when used to compute tweet-level polarity detection with context-based adaptation. While enriching the adaptation process with words' semantic subtypes has modest impact on the lexicons' performance, Enrichment based on the semantic relations between entities in tweets, yields in 4.12% and 3.12% gain in accuracy and F1 measure in comparison with context-based adaptation. Our results also showed that the lexicons adapted using our proposed method are more robust to imbalanced datasets.

[8]The effective estimated diameter of DBPedia is 6.5082 edges. See `http://konect.uni-koblenz.de/networks/dbpedia-all`

Acknowledgment

This work was supported by the EU-FP7 project SENSE4US (grant no. 611242).

Appendix A

Tables 11 and 12 lists the complete results of using the conceptually-enriched model and the semantically-adjusted relations models on the three evaluation datasets respectively.

References

[1] Amir Asiaee T, Mariano Tepper, Arindam Banerjee, and Guillermo Sapiro. If you are happy and you know it... tweet. In *Proc. 21st ACM conference on Information and knowledge management*, 2012.

[2] S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Seventh conference on International Language Resources and Evaluation.*, Valletta, Malta, 2010.

[3] Carmen Banea, Rada Mihalcea, and Janyce Wiebe. A bootstrapping method for building subjectivity lexicons for languages with scarce resources. In *LREC*, volume 8, pages 2–764, 2008.

[4] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[5] Erik Cambria. An introduction to concept-level sentiment analysis. In *Advances in Soft Computing and Its Applications*, pages 478–483. Springer, 2013.

[6] Ilia Chetviorkin, Leninskiye Gory Moscow, and Natalia Loukachevitch. Two-step model for sentiment lexicon extraction from twitter streams. *ACL 2014*, page 67, 2014.

[7] Yejin Choi and Claire Cardie. Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. In *Proc. Conf. Empirical Methods in Natural Language Processing*, 2009.

[8] Weifu Du, Songbo Tan, Xueqi Cheng, and Xiaochun Yun. Adapting information bottleneck method for automatic construction of domain-oriented sentiment lexicon. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 111–120. ACM, 2010.

[9] Shi Feng, Kaisong Song, Daling Wang, and Ge Yu. A word-emoticon mutual reinforcement ranking model for building sentiment lexicon from massive collection of microblogs. *World Wide Web*, pages 1–19, 2014.

[10] Zellig S Harris. Distributional structure. *Word*, 1954.

Dataset	Lexicon	Accuracy	Negative Sentiment			Positive Sentiment			Average		
			P	R	F1	P	R	F1	P	R	F1
SemEval	Original	71.85	50.84	85.17	63.67	91.66	66.42	77.02	71.25	75.79	70.35
	SO-PMI	73.9	53.19	82.37	64.64	90.74	70.44	79.31	71.96	76.41	71.97
	SLU	76.12	56.2	79.24	65.72	89.84	74.84	81.65	73.02	77.04	73.68
	SLE	72.95	52.17	79.71	63.03	89.45	70.2	78.65	70.81	74.96	70.84
	SLUE	**76.18**	56.31	79.35	65.83	89.89	74.9	81.7	**73.1**	**77.12**	**73.76**
WAB	Original	**79.24**	77.96	77.84	77.9	80.37	80.48	80.43	79.17	**79.16**	**79.16**
	SO-PMI	79.04	80.49	73.15	76.64	77.96	84.27	80.99	79.22	78.71	78.82
	SLU	79.13	81.13	72.45	76.52	77.68	85.06	81.18	**79.41**	78.76	78.85
	SLE	79.17	78.13	77.34	77.71	80.07	80.78	*80.4*	79.1	79.06	79.05
	SLUE	79.04	81.05	72.34	76.41	77.59	84.99	81.1	*79.32*	78.66	78.75
GASP	Original	69.38	86.89	74.25	80.08	26.88	45.79	33.87	56.88	60.02	56.97
	SO-PMI	69.69	86.73	74.89	80.38	26.82	44.53	33.48	56.77	59.71	56.93
	SLU	**73.04**	87.48	78.73	82.87	30.68	45.54	36.57	**59.08**	**62.14**	**59.72**
	SLE	69.21	86.9	74.02	79.93	26.78	46	33.77	56.84	60.01	56.85
	SLUE	73.03	87.47	78.74	82.86	30.62	45.41	36.5	59.05	62.07	59.68
Average	Original	73.49	71.90	79.09	73.88	66.30	64.23	63.77	69.10	71.66	68.83
	S0-PMI	74.21	73.47	76.80	73.89	65.17	66.41	64.59	69.32	71.61	69.24
	SLU	**76.47**	75.54	76.31	75.29	65.93	69.20	66.87	**70.74**	**72.75**	**71.08**
	SLE	73.78	72.40	77.02	73.58	65.44	65.66	*64.31*	68.92	71.34	68.94
	SLUE	76.42	75.47	76.31	75.24	65.90	*69.10*	66.81	*70.69*	72.71	71.03

Table 11

Results obtained from adapting Thelwall-Lexicon on three datasets using the conceptually-enriched adaptation model. Bold=highest performance. Italic=significance at 0.05, None-Italic=significance $<$ 0.001, *=significance at 0.1

Dataset	Lexicon	Accuracy	Negative Sentiment			Positive Sentiment			Average		
			P	R	F1	P	R	F1	P	R	F1
SemEval	Original	71.85	50.84	85.17	63.67	91.66	66.42	77.02	71.25	75.79	70.35
	SO-PMI	73.9	53.19	82.37	64.64	90.74	70.44	79.31	71.96	76.41	71.97
	SRU	75.66	55.8	76.76	64.59	88.81	75.22	81.44	72.31	75.99	73.01
	SRE	**77.23**	58.27	75.34	65.67	88.58	78	82.94	**73.43**	**76.67***	**74.31**
	SRUE	75.72	55.88	76.8	64.66	88.85	75.27	81.48	72.36	76.04	73.07
WAB	Original	**79.24**	77.96	77.84	77.9	80.37	80.48	80.43	79.17	**79.16**	**79.16**
	SO-PMI	79.04	80.49	73.15	76.64	77.96	84.27	80.99	79.22	78.71	78.82
	SRU	78.97	84.17	68.06	75.23	75.78	88.64	81.68	79.97	78.35	78.46
	SRE	78.95	84.8	67.31	75.01	75.47	89.29	81.78	**80.14**	78.3	78.39
	SRUE	78.89	84.11	67.98	75.15	75.72	88.56	81.62	79.91	78.27	78.39
GASP	Original	69.38	86.89	74.25	80.08	26.88	45.79	33.87	56.88	*60.02*	56.97
	SO-PMI	69.69	86.73	74.89	80.38	26.82	44.53	33.48	56.77	59.71	56.93
	SRU	**72.96**	86.81	79.44	82.95	29.51	41.58	34.43	58.16	60.51	58.69
	SRE	72.8	87.41	78.47	82.69	30.25	45.3	36.22	**58.83**	**61.89**	**59.46**
	SRUE	72.91	86.82	79.35	82.9	29.43	41.68	34.42	58.12	60.52	58.66
Average	Original	73.49	71.90	79.09	73.88	66.30	64.23	63.77	69.10	71.66	68.83
	S0-PMI	74.21	73.47	76.80	73.89	65.17	66.41	64.59	69.32	71.61	69.24
	SRU	76.14	75.99	74.74	74.50	64.78	68.95	66.14	70.39	71.84	70.31
	SRE	**76.66**	77.38	73.62	74.76	64.84	71.42	67.31	**71.11**	**72.52**	**71.03**
	SRUE	76.13	76.03	74.75	74.52	64.78	69.01	66.16	70.41	71.88	70.34

Table 12

Results obtained from adapting Thelwall-Lexicon on three datasets using the semantically-adjusted relations model. Bold=highest performance. Italic=significance at 0.05, None-Italic=significance $<$ 0.001, *=significance at 0.1

[11] Xia Hu, Jiliang Tang, Huiji Gao, and Huan Liu. Unsupervised sentiment analysis with emotional signals. In *Proceedings of the 22nd World Wide Web conf*, 2013.

[12] Yongyos Kaewpitakkun, Kiyoaki Shirai, and Masnizah Mohd. Sentiment lexicon interpolation and polarity estimation of objective and out-of-vocabulary words to improve sentiment classification on microblogging. In *Proc. 28th Pacific Asia Conf. Language, Information and Computing*, Thailand, 2014. Department of Linguistics, Faculty of Arts, Chulalongkorn University.

[13] Hiroshi Kanayama and Tetsuya Nasukawa. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 355–363. Association for Computational Linguistics, 2006.

[14] Fangtao Li, Sinno Jialin Pan, Ou Jin, Qiang Yang, and Xiaoyan Zhu. Cross-domain co-extraction of sentiment and topic lexicons. In *Proc. 50th Annual Meeting of Association for Computational Linguistics*, 2012.

[15] Chenghua Lin, Yulan He, Richard Everson, and Stefan Ruger. Weakly supervised joint sentiment-topic detection from text. *Knowledge and Data Engineering, IEEE Transactions on*, 24(6):1134–1145, 2012.

[16] Yue Lu, Malu Castellanos, Umeshwar Dayal, and ChengXiang Zhai. Automatic construction of a context-aware sentiment lexicon: an optimization approach. In *Proc. Int. Conf. World Wide Web*, 2011.

[17] Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. Semeval-2013 task 2: Sentiment analysis in twitter. In *Proc. 7th ACL Workshop on Semantic Evaluation.*, 2013.

[18] Finn Årup Nielsen. A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*, 2011.

[19] Giuseppe Pirrò. Explaining and suggesting relatedness in knowledge graphs. In *The Semantic Web-ISWC 2015*, pages 622–639. Springer, 2015.

[20] Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Expanding domain sentiment lexicon through double propagation. In *IJCAI*, volume 9, pages 1199–1204, 2009.

[21] Hassan Saif, Miriam Fernandez, Yulan He, and Harith Alani. Evaluation datasets for twitter sentiment analysis. In *Proceedings, 1st ESSEM Workshop*, Turin, Italy, 2013.

[22] Hassan Saif, Miriam Fernandez, Yulan He, and Harith Alani. SentiCircles for contextual and conceptual semantic sentiment analysis of twitter. In *Proc. 11th Extended Semantic Web Conf. (ESWC)*, Crete, Greece, 2014.

[23] Hassan Saif, Yulan He, and Harith Alani. Semantic sentiment analysis of twitter. In *Proc. 11th Int. Semantic Web Conf. (ISWC)*, Boston, MA, 2012.

[24] Hassan Saif, Yulan He, Miriam Fernandez, and Harith Alani. Adapting sentiment lexicons using contextual semantics for sentiment analysis of twitter. In *ESWC 2014 Satellite Events*. 2014.

[25] Hassan Saif, Yulan He, Miriam Fernandez, and Harith Alani. Contextual semantics for sentiment analysis of twitter. *Information Processing & Management*, 2015.

[26] Hassan Saif, F Javier Ortega, Miriam Fernández, and Iván Cantador. Sentiment analysis in social streams.

[27] Christian Scheible and Hinrich Schütze. Bootstrapping sentiment labels for unannotated documents with polarity pagerank. In *LREC*, pages 1230–1234, 2012.

[28] Sidney Siegel. Nonparametric statistics for the behavioral sciences. 1956.

[29] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment strength detection for the social web. *J. American Society for Information Science and Technology*, 63(1):163–173, 2012.

[30] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment strength detection in short informal text. *J. American Society for Info. Science and Technology*, 61(12), 2010.

[31] P. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, Philadelphia, Pennsylvania, 2002.

[32] Peter Turney and Michael Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21:315–346, 2003.

[33] Peter D Turney, Patrick Pantel, et al. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188, 2010.

[34] Warren Weaver. Translation. *Machine translation of languages*, 14:15–23, 1955.