Selective Co-occurrences for Word-Emotion Association

Ameeta Agrawal and Aijun An
Department of Electrical Engineering and Computer Science
York University, Toronto, Canada
{ameeta,aan}@cse.yorku.ca

Abstract

Emotion classification from text typically requires some degree of word-emotion association, either gathered from pre-existing emotion lexicons or calculated using some measure of semantic relatedness. Most emotion lexicons contain a fixed number of emotion categories and provide a rather limited coverage. Current measures of computing semantic relatedness, on the other hand, do not adapt well to the specific task of word-emotion association and therefore, yield average results. In this work, we propose an unsupervised method of learning word-emotion association from large text corpora, called **Se**lective **Co**-occurrences (SECO), by leveraging the property of mutual exclusivity generally exhibited by emotions. Extensive evaluation, using just one seed word per emotion category, indicates the effectiveness of the proposed approach over three emotion lexicons and two state-of-the-art models of word embeddings on three datasets from different domains.

1 Introduction

Emotion detection from text is the task of identifying emotions from natural language data such as user reviews, blogs, news articles, etc. (Alm et al., 2005; Aman and Szpakowicz, 2007). Although there is no strict definition for emotion, most researchers agree that it is a particular feeling that characterizes the state of mind such as happiness, anger, sadness and so on. Emotion analysis is extremely popular in the field of market research, where brands tap into their customer base by analyzing user-generated data, readily available nowadays due to the rapid growth of social media (De Bondt et al., 2013; Shrum et al., 2013). Mohammad and Turney (2013) present a more extensive list of applications.

Emotion detection from text typically requires some degree of *word-emotion association*, i.e., knowing which words are more appropriately associated with which emotions. For example, the word *"accident"* can be considered associated with the emotion SADNESS. This association can be obtained from a pre-compiled emotion lexicon or calculated using some measure of semantic relatedness. The currently available manually annotated emotion lexicons tend to be restricted in size due to the expensive process of human annotation. This limited coverage leads to the undesirable effect of leaving too many words unassociated with any emotion category. Moreover, most lexicons contain a small fixed set of emotions which is unsuitable for a larger (Du et al., 2014) or a newly defined set of emotions (Facebook, 2016). On the other hand, while automatically computing word-emotion association scores from text corpora possibly provides a better coverage and more flexibility, the current techniques are ill-suited to the task of emotion detection and therefore, tend to yield average results.

To address these issues, in this work, we propose an unsupervised method of learning word-emotion association scores from text corpora, which we call **Se**lective **Co**-occcurences (SECO). By modifying conventional co-occurrence-based methods, we compute a uni-directional asymmetric association between a given word and an emotion seed word. The proposed approach is found to be better at capturing the association between words and emotions than general purpose measures. Extensive evaluation of word-emotion association scores derived from two large text corpora, Wikipedia and Amazon reviews,

on three emotion datasets from very diverse domains demonstrates the effectiveness of employing selective co-occurrences. The proposed approach is particularly interesting as it requires no training data and can be applied to a flexible number of emotion categories.

The remainder of this paper is organized as follows. In Section 2, we survey the related work. In Section 3, we describe the learning of the word-emotion association scores and their use in the task of emotion classification. Section 4 describes the evaluation setup, while Section 5 analyzes the experimental results. Lastly, Section 6 concludes the paper with a brief look at future work.

2 Related Work

In this section, we present the existing emotion lexicons, manually annotated as well as those created using supervised machine learning, and also discuss the measures of semantic relatedness that have been previously employed to derive word-emotion association for emotion classification.

2.1 Emotion Lexicons

2.1.1 Manually Annotated Lexicons

One of the earliest and most popular emotion resources is the WordNet Affect (Strapparava and Valitutti, 2004), developed by manually labelling about 1,314 synsets with one or more of Ekman's (1992) six basic emotions. Using crowd-sourcing, one of the largest manually annotated emotion lexicons created to date is the NRC Emotion Lexicon (EmoLex) (Mohammad and Turney, 2010; Mohammad and Turney, 2013). It contains about 14,200 unigrams annotated with one or more of Plutchik's eight emotions (Plutchik, 2001). Another manually created lexicon, the Affect database (Neviarouskaya et al., 2007), contains a total of 2,440 entries (emoticons, acronyms, words and modifiers) annotated by three annotators using nine emotion labels as well as their intensities. Considering the fundamental role played by lexical resources in the task of emotion detection, the current options available due to manual lexicons seem rather limited in their coverage. Human annotation, including crowd-sourcing, requires considerable lexicographic expertise, time and effort. An alternative approach involves creating such lexical resources automatically.

2.1.2 Automatically Acquired Lexicons

More recently, DepecheMood (Staiano and Guerini, 2014), was created using supervised training by applying distributional semantics to a dataset of crowd-annotated news articles. This lexicon consists of 37,000 words and their emotion scores across seven emotions. A different approach of generating an emotion lexicon (18,000 words and 8 emotions) involves using the Google n-grams corpus to expand an existing, smaller human-annotated lexicon such as the EmoLex (Perrie et al., 2013). Although these approaches cover a larger vocabulary, they are limited to the emotion categories of the source corpus or the lexicon which they use for training.

2.2 Measures of Semantic Relatedness

Statistical approaches that leverage large text corpora provide an alternative way of acquiring word-emotion association scores, which can remedy the problem of unseen vocabulary to a large extent. As these approaches employ just a handful of emotion seed words to initialize the process, they are also applicable to a flexible number of emotion categories. Many models of computing word semantic relatedness exist, from the traditional count-based methods such as Pointwise Mutual Information (PMI) to the more recent neural-network inspired models of word embeddings such as Continuous Bag of Words (CBOW). While the models differ in their algorithms, they are fundamentally based on the intuitive assumption that co-occurring words tend to be related to each other.

Previously, PMI has been used to classify emotions in news headlines, where the probabilities of words were calculated using statistics collected from three search engines (Kozareva et al., 2007). Wikipedia has also proven a useful resource for calculating word frequencies using PMI to obtain word-emotion association scores (Agrawal and An, 2012). Alternatively, PMI has been used to first build an emotion lexicon which is then used for classification (Yang et al., 2007). Latent Semantic Analysis (LSA)

(Deerwester et al., 1990), which analyzes the statistical relationships among words in a corpus using Singular Value Decomposition (SVD) dimensionality reduction technique, was used to calculate word-emotion association scores to classify news headlines (Strapparava and Mihalcea, 2008).

Despite, and perhaps because of, its simplicity, PMI (Church and Hanks, 1990) has long been a popular measure of semantic relatedness. It estimates the similarity between two terms x and y as $PMI\,(x,y) = \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$, where $p\,(x,y)$ is the probability that words x and y co-occur within a window of specific length, and $p\,(x)$ and $p\,(y)$ are the individual probabilities of word x and word y, respectively, in the corpus. To overcome a few well-known shortcomings of PMI (i.e., low frequency events receiving relatively high scores, lack of a fixed upper bound), Bouma (2009) proposed a normalized version of PMI (NPMI), where $NPMI\,(x,y) = \frac{PMI(x,y)}{-\log p(x,y)}$, with fixed orientation values: when two words only occur together, NPMI$(x,y) = 1$; when they are distributed as expected under independence, NPMI$(x,y) = 0$; and, when they occur separately but not together, NPMI$(x,y) = -1$.

More recently proposed neural-network based approaches, such as Continuous Bag-of-Words (CBOW) and Skip-Gram (SG) (Mikolov et al., 2013b; Mikolov et al., 2013a) output word embeddings which can then be used to compute the similarity between two words by calculating their cosine similarity. These methods implicitly factorize a word-context matrix whose cell values are in fact shifted PMI (Levy and Goldberg, 2014) and have outperformed traditional count-based methods such as PMI on many similarity-oriented tasks (Marco Baroni, Georgiana Dinu, 2014).

While these measures of semantic relatedness provide promising results on most word similarity tasks, they do not adapt well to emotion detection. We build upon these latest advancements by proposing a variant that is more suitable to the task at hand.

3 Word-emotion Association for Emotion Classification

Given a sentence s and a set $E = \{e_1, e_2, \ldots, e_g\}$ of g emotion categories, the objective is to label s with the best possible emotion $e \in E$. We first discuss our proposed method for deriving the word-emotion association scores between a word and an emotion category, and then use these scores to obtain an emotion label for each sentence.

3.1 Learning Word-emotion Association

Let $W = \{w_1, w_2, \ldots, w_n\}$ be a set of n cue words in an input sentence s, where $W \subset s$. A cue word is defined as any word within a sentence that could have some emotional connotation. Usually, these are the nouns, adjectives, verbs and adverbs. Let $E = \{e_1, e_2, \ldots, e_g\}$ be a set of g emotion categories. Each emotion category $e_j \subset E$ is represented by one or more emotion seed words. Let T be the set of all the seed words for all the emotion categories, $T_j \subset \mathrm{T}$ and $T_j = \{t_{j_1}, t_{j_2}, \ldots, t_{j_m}\}$ be the set of m emotion seed words for each emotion category $e_j \subset E$.

As an illustration, consider the sentence "*We are going to a party tonight*". Here, the set of cue words includes $W = \{going, party, tonight\}$. If the classification scheme follows, for example, Ekman's (1992) model of emotions, then $E = \{anger, disgust, fear, happiness, sadness, surprise\}$. The set of seed words for HAPPINESS could be $T_{happiness} = \{happy, joy, ...\}$ and ANGER's seed words could be $T_{anger} = \{angry, mad, \ldots\}$.

We adopt the $\rightarrow$ symbol to denote the association between a cue word w and an emotion seed word t. The first step is to derive the association scores between a cue word and every seed word, e.g., $Assoc.\,(party \rightarrow joy)$, $Assoc.\,(party \rightarrow angry)$. Since our main goal is to acquire *association scores for emotion classification*, the design choices for our proposed measure of learning word-emotion association scores, SECO, are largely motivated by the following observations:

- The intrinsic process of annotating an emotion dataset as well as that of classifying it is uni-directional, i.e., given a word or a sentence, the task is to label it with the emotion it evokes the most.
- Although expressions of emotions can sometimes be fuzzy, most words primarily evoke only one emotion in a particular context, i.e., the emotion categories are, for the most part, mutually exclusive.

In fact, in the emotion lexicon WordNet Affect (Strapparava and Valitutti, 2004), which has words annotated with more than one emotion, about 98.7% of the terms are labelled with just one emotion.

- Most importantly, unlike other word relatedness tasks, the second half of the association pair (i.e., emotion seed words) in this particular task are known in advance.

These observations lead us to hypothesize that an asymmetric measure of association, where a word's association with an emotion seed word (and therefore by extension, with an emotion category) may be more meaningful than trying to achieve a symmetric bi-directional association between the two. In fact, as Tversky (1977) notes, certain linguistic relationships are characteristically asymmetric . In one experiment to list the first meaningfully related word that comes to mind, for the cue word *fear*, 24% of the participants answered *scared*, while only 9% of them recalled *fear* when given the cue word *scared*, suggesting an inherent asymmetry in word associations (Altarriba et al., 1999).

As noted earlier, traditional co-occurrence based models of semantic relatedness consider two words as co-occurring when both the words appear within a specific window of text, no matter how far they are from each other. In reality, nearer words have been found to exhibit stronger relationships (Beeferman et al., 1997).

In similar essence, emotions generally exhibit the property of mutual exclusivity and therefore, we propose the concept of selective co-occurrences (SECO), where a cue word is considered as co-occurring with only one emotion's seed words within any particular window of text. Consider Fig. 1 containing an example window of text, the cue word "*party*", and two seed words "*angry*" and "*happy*", representing the two different emotion categories respectively. To apply selective co-occurrences, the cue word "*party*" is considered co-occurring with either "*angry*" or "*happy*", not both.

> Theater critic Michael Riedel (playing himself) also shows up, uninvited. Ivy is put out by this and gets **angry** at Michael about it. We hear but don't see Ivy singing "Bittersweet Symphony" at her **party**. Derek then walks in and gives her a present and wishes her **happy** birthday.

Figure 1: Sample window of text containing cue and seed words

When a context window contains multiple seed words from multiple emotion categories, three possible settings for selecting the most appropriate seed word as co-occurring with the cue word can be explored:

- *nearest* (SECO-NEAR): This is the most intuitive option, where the nearest seed word to the cue word is selected. For example, "*happy*" is counted as co-occurring with "*party*"; "*angry*" is ignored.
- *preceding* (SECO-PREC): To account for any positional predisposition, this variant considers only the closest preceding seed word to the cue word. For example, "*angry*" is considered as co-occurring with "*party*"; "*happy*" is ignored.
- *following* (SECO-FOLL): Similarly, this variant considers only the closest seed word that follows the cue word as co-occurring together. For example, "*happy*" is considered as co-occurring with "*party*"; "*angry*" is ignored.

The selective counting of the seed word's co-occurrence frequency with the cue word is what essentially makes our association measure asymmetric and therefore, the order of the cue and seed word in the association equation cannot be interchanged. That is, $Assoc.(w \rightarrow t)$ denotes the association between a cue word w and a seed word t, and $Assoc.(w \rightarrow t) \neq Assoc.(t \rightarrow w)$. Technically, SECO is applicable to any traditional co-occurrence-based word association measure that estimates the relatedness between two words by computing some function of the words' frequencies. To this end, we adopt three popular co-occurrence association measures, namely NPMI (Bouma, 2009), Dice (1945) and Jaccard (1912), into their SECO counterparts, SECO-NPMI, SECO-Dice and SECO-Jaccard, respectively.

When applying SECO, in a corpus of M words, $\#(w, t)$ denotes the number of times a cue word w co-occurs mutually exclusively with any one emotion seed word $t \subset \mathrm{T}$ within a context window of size

k, where k is the maximum distance between those two words; $\#(w)$ and $\#(t)$ denote the frequencies of w and t, respectively.

SECO-NPMI The normalized SECO-NPMI between w and t, within the range of $[-1, 1]$, is:

$$SECO\text{-}NPMI(w \rightarrow t) = \frac{\log\left(M \frac{\#(w,t)}{\#(w)\#(t)}\right)}{\log\left(\frac{M}{\#(w,t)}\right)}. \quad (1)$$

SECO-Dice Similarly, SECO-Dice between w and t is computed as:

$$SECO\text{-}Dice(w \rightarrow t) = \frac{2 \times \#(w,t)}{\#(w) + \#(t)}. \quad (2)$$

SECO-Jaccard Lastly, one of the earliest co-occurrence associations measures, Jaccard, can be transformed as follows:

$$SECO\text{-}Jaccard(w \rightarrow t) = \frac{\#(w,t)}{\#(w) + \#(t) - \#(w,t)}. \quad (3)$$

Empirical data shows that the association between words decays exponentially (Beeferman et al., 1997) and this property has been successfully exploited by adding a decaying factor which allows words that co-occur nearer to each other to be more related (Gao et al., 2002; Sahlgren, 2006; Brosseau-Villeneuve et al., 2010; Mikolov et al., 2013a). Drawing from previous research, we also apply a context weighting scheme whereby a seed word is linearly weighted according to its distance from the cue word as follows: in a window of size k, the n^{th} word from the cue word is weighted by the function $\frac{k-n+1}{k}$. For example, in a window of 5, the first word next to the cue word is weighted by $\frac{5}{5}$, while the fourth word away is of weight $\frac{2}{5}$. In other words, as the distance between two words increases, their weighted association score decreases.

Finally, the word-emotion association between w and an emotion category e_j is obtained by calculating the average mean of the association scores between w and all the seed words of e_j as:

$$Assoc.(w \rightarrow e_j) = \frac{1}{m} \sum_{k=1}^{m} Assoc.(w \rightarrow t_{j_k}) \quad (4)$$

where $Assoc.$ is any association measure such as SECO-NPMI.

3.2 Classifying Sentence Emotion

For each word w, its emotion vector ϕ_w is denoted as:

$$\phi_w = \langle Assoc.(w \rightarrow e_1), Assoc.(w \rightarrow e_2), \ldots, Assoc.(w \rightarrow e_g) \rangle$$

and the emotion vector ϕ_s of sentence s is obtained by averaging the emotion vectors of all its n cue words as $\phi_s = \frac{1}{n} \sum_{i=1}^{n} \phi_{w_i}$. Finally, the sentence is labelled with the emotion category $e \subset E$ with the maximum value in ϕ_s.

4 Evaluation Setup

In this section we describe the evaluation datasets, the text corpus used for learning the word-emotion association scores and the evaluation metric for this task.

4.1 Evaluation Datasets

Below described are the three popular emotion evaluation datasets, with some sample sentences presented in Table 1 and their summarized statistics in Table 2.

Aman: Consisting of highly informal blog data, this dataset includes 1290 sentences annotated with one of six emotions: *anger, disgust, fear, joy, sadness and surprise* (Aman and Szpakowicz, 2007).

Alm: Emotions are particularly significant in the literary genre of fairy tales and this dataset contains 1207 high-agreement sentences (i.e., all four annotators agreed with the same label) marked with one of five emotions: *angry-disgusted, fearful, happy, sad and surprised* (Alm, 2008).

Aman	Do you people not listen to the news or what? I had a blast in california hanging out with my family and friends.	*anger* *happiness*
Alm	Oh! cried the devil, "what are you doing?" Ha! what are you doing? cried the devil angrily.	*surprised* *angry-disgusted*
ISEAR	When I saw a ghost. Slaughtering of animals.	*fear* *disgust*

Table 1: Sample sentences from evaluation datasets

	ag	***dg***	***fr***	***hp***	***sd***	***sp***	**total**
Aman	179	172	115	536	173	115	1290
Alm	218		166	445	264	114	1207
ISEAR	1085	1072	1086	1089	1080	-	5412

Table 2: Details of evaluation datasets

ISEAR: Developed for studying the relationships among emotions and cultures, this corpus contains experiences evoking seven emotions: *anger, disgust, fear, joy, sadness, shame* and *guilt*, resulting in a total of 5412 sentences[1]. To the best of our knowledge, no existing lexicon contains *shame* or *guilt* categories, and therefore methods that depend on emotion lexicons cannot correctly classify sentences belonging to these emotions. However, unsupervised approaches such as ours, which can be initialized with as little as one seed word per emotion category, are easily applicable to such datasets.

4.2 Text Corpora

We derive the word-emotion association scores from two large text corpora of different domains, which are pre-processed by: a) converting to lowercase; b) stripping off all non-alphanumeric characters; c) removing stopwords; d) stemming, and e) removing words that occur less than 5 times in the corpus.

Wikipedia[2]: The large publicly available corpus of Wikipedia mainly consists of formal language structured text articles considered to be more "objective" in nature. Our clean corpus contains approximately 918.5 million tokens, with each article on one line.

Amazon: The text of all the product reviews, mostly consisting of informal language makes up our second corpus, considered to be of more "emotional" type. This data was extracted from the aggressively deduplicated dataset (McAuley et al., 2015), which contains 82.83 million product reviews from Amazon, spanning May 1996 - July 2014. Our clean corpus contains more than 3 billion tokens (three times the size of Wikipedia corpus), with one review per line.

4.3 Evaluation Metric

Following prior studies, we calculate the F-score for each emotion class e, where F-score is the harmonic mean of *precision* and *recall*, defined as $2\left(\frac{precision \times recall}{precision + recall}\right)$. *Precision* is the number of sentences correctly labeled as belonging to the class e divided by the total number of sentences labeled as belonging to e, and *recall* is the number of sentences correctly labeled as belonging to e divided by the total number of sentences that actually belong to e. We report the average F-score over all the classes.

5 Experiments and Analysis

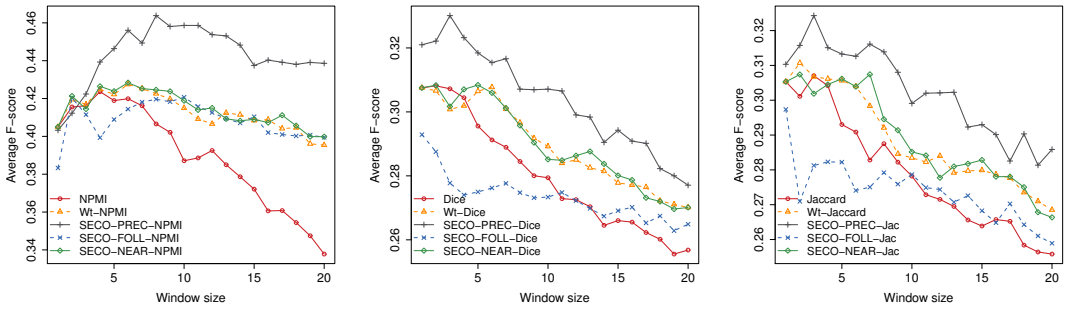
In what follows, we evaluate the performance of the proposed approach in several experiments and discuss their results.

5.1 How effective is selective co-occurrence?

We test the performance of the three proposed variants, `SECO-NEAR`, `SECO-PREC` and `SECO-FOLL`, against regular as well as weighted versions (where the same context weighting scheme as described in Section 3.1 is applied to regular association measures) in order to analyze the effect of selective co-occurrence in particular, and not just the advantage obtained using weighted contexts.

[1] http://www.affective-sciences.org/system/files/webpage/ISEAR_0.zip. Removed instances with [no response].

[2] http://dumps.wikimedia.org/enwiki/20140811/enwiki-20140811-pages-articles.xml.bz2



(a) NPMI (b) Dice (c) Jaccard

Figure 2: Selective co-occurrences and regular co-occurrences on Alm dataset

As observed from Figure 2, tested for context window sizes 1 to 20 and trained on the Wikipedia corpus, our proposed approach, *preceding* selective co-occurrences (SECO−PREC), where only the nearest seed word that precedes a given cue word within a context window is considered, exhibits the best performance when applied to all the three association measures (NPMI, Dice and Jaccard), on the Alm dataset[3]. In fact, the best average F-score from SECO−NPMI−PREC is almost 10% better than that of Wt-NPMI, leading us to conclude that the gain in performance is due to selective co-occurrences and not just weighted contexts. As expected, the weighted versions, Wt-NPMI, Wt-Dice and Wt-Jaccard perform much better than their regular unweighted counterparts, NPMI, Dice and Jaccard, respectively. SECO−NEAR has slight advantage over weighted association measures while SECO−FOLL and regular association measures (i.e., NPMI, Dice, Jaccard) are nearly always the poorest performing. Since SECO−PREC−NPMI (Figure 2a) yielded better average F-score than SECO−PREC−Dice or SECO−PREC−Jaccard, we further evaluate its performance in the next experiment.

5.2 How effective is SECO−PREC−NPMI?

In this experiment, we evaluate the performance of unsupervised SECO−PREC−NPMI against five baselines (in **bold**) described next.

5.2.1 Baselines

WordNet Affect (**WNA**), NRC Emotion Lexicon (**EmoLex**) and DepecheMood (**DM**) (Section 2.1) contain words and their association with various emotions. For each emotion category, WNA contains a simple list of words, which we interpret as a binary association; if a word exists in an emotion category, we assign $+1$ for that emotion. EmoLex and DM, on the other hand, contain association scores between a word and all the emotion categories. For instance, EmoLex lists the association between the word "*awful*" and 8 emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust) as: $1, 0, 1, 1, 0, 1, 0, 0$. In DM, each word is associated with a different set of emotions by a real valued score, summing upto 1. For instance, the word "*awe*" and 8 emotions (afraid, amused, angry, annoyed, dont_care, happy, inspired, sad) is listed as: $0.08, 0.12, 0.04, 0.11, 0.07, 0.15, 0.38, 0.05$. We use these emotion lexicons as a baseline by applying a keyword matching algorithm, to obtain $Assoc.(w \rightarrow e)$. For example, $Assoc.(awe \rightarrow sad) = 0.05$ from DM. Note that, since DM does not contain two of the emotions found in our evaluation datasets, i.e., disgust and surprise, we report its results on a subset of the datasets. Instead of directly comparing our word-emotion association scores with those of emotion lexicons, we extrinsically evaluate them in the task of emotion classification as there are significant differences between the various emotion lexicons and none can be considered to be a perfect benchmark.

Semantic similarity computed from two state-of-the-art word embedding algorithms, **CBOW** and **SG** (Mikolov et al., 2013b; Mikolov et al., 2013a) (Section 2.2), provide another baseline as they can be used to obtain unsupervised word-emotion association scores, which is closer in spirit to our goal. We

[3]Consistent results were obtained on Aman and ISEAR datasets, not included here due to limited space.

used the algorithms' recommended default parameter settings: *dimension size of feature vectors* = 300; *negative sampling* = 5. We present a comparison against only unsupervised methods as the focus of this paper is on unsupervised emotion detection methods.

Since the context window size can have a significant impact on the performance of an algorithm, we run each method of semantic relatedness on 20 different window sizes (1 to 20) on both the corpora (Wikipedia and Amazon) and report the average result with standard deviation for each setting in Table 3. To keep the process as unsupervised as possible, in this study only one seed word per emotion category is used to derive the association scores. The seed words "*angry, disgust, happy, scared, sad, surprise*" represent the six emotion categories "*anger, disgust, happiness, fear, sadness, surprise*", respectively.

	Aman	**Alm**	**ISEAR**
SG_{wiki}	0.242 ± 0.04	0.209 ± 0.05	0.259 ± 0.08
$CBOW_{wiki}$	0.382 ± 0.02	0.426 ± 0.03	0.446 ± 0.03
SECO-PREC-NPMI$_{wiki}$	**0.410** ± 0.01**	**0.443** ± 0.01**	0.488 ± 0.02**
SG_{amazon}	**0.410** ± 0.02*	0.406 ± 0.02	0.438 ± 0.04
$CBOW_{amazon}$	0.393 ± 0.02	0.373 ± 0.02	0.484 ± 0.03
SECO-PREC-NPMI$_{amazon}$	0.403 ± 0.01	0.409 ± 0.01**	**0.498** ± 0.02**

Table 3: Average F-scores (of windows 1 to 20) for three evaluation datasets. SG, CBOW and SECO-PREC-NPMI were run on Wikipedia and Amazon corpora for windows 1 to 20. The best average result for each dataset is in **bold**. $^{**}p < .00001$, $^{*}p < .01$ (one-way ANOVA test for each dataset results using the same training corpus, i.e., wiki or amazon)

5.2.2 Results

Usually it is difficult to determine the best window size in advance, and therefore, for window sizes 1 to 20, we summarize the average F-scores over all 20 window sizes in Table 3. The best results obtained using any particular window are presented in Table 4 and the details of different emotion category results are further shown in Table 5. Finally, Figure 3 presents the window sensitivity graphs.

i) The average F-score results summarized in Table 3 indicate that on average, SECO-PREC-NPMI yields better overall results, with SG_{amazon}obtaining competitive results on one dataset (Aman), suggesting the effectiveness of selective co-occurrences in this task. Interestingly, contrary to popular intuition, the "objective" text from Wikipedia training corpus yields better F-scores on average than the "subjective" Amazon reviews corpus for two out of the three evaluation datasets.

	Aman	**Alm**	**ISEAR**
WNA	0.286	0.362	0.343
EmoLex	0.316	0.341	0.318
DM	0.324	0.340	0.290
SG_{wiki}	0.338 (2)	0.345 (1)	0.433 (1)
$CBOW_{wiki}$	0.410 (15)	0.456 (11)	0.481 (19)
SECO-PREC-NPMI$_{wiki}$	0.422 (10)	**0.464** (8)	0.497 (10)
SG_{amazon}	**0.435** (6)	0.440 (1)	0.490 (5)
$CBOW_{amazon}$	0.411 (6)	0.399 (18)	0.510 (19)
SECO-PREC-NPMI$_{amazon}$	0.412 (11)	0.422 (15)	**0.512** (20)

Table 4: Details of best results for three evaluation datasets. The best result for each dataset is in **bold**. The window size is shown in parentheses.

ii) The results of Table 4 indicate that, on all three evaluation datasets, with just one seed word per emotion category used to derive the word-emotion association scores, all the unsupervised measures of semantic relatedness (SECO-PREC-NPMI, SG and CBOW) outperform all the emotion lexicons (WNA, EmoLex and DM) that were created using considerable human input and training data, indicating that semantic similarity approaches provide an effective unsupervised way of extracting meaningful word-emotion association scores. Within the emotion lexicons, WNA provides the best performance on two out

of the three datasets despite being the smallest in size. Unsupervised association measures demonstrate two significant advantages over emotion lexicons: firstly, association measures are able to provide a wider coverage by exploiting the inherent associations between words that are present in text corpora; and secondly, while lexicons have fixed pre-determined emotion categories, association measures can be flexibly extended to any number and types of emotions. As for the recommended window settings for each approach, it seems that SG works well with window size $= 1$ on Wikipedia corpus and around 5 on Amazon; CBOW usually does well on windows larger than 15 and `SECO-PREC-NPMI` is recommended to be used with window 10 on Wikipedia and larger than 15 on Amazon.

AMAN																		
	ag			*dg*			*fr*			*hp*			*sd*			*sp*			Avg
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	
SG_{amazon}	0.66	0.25	0.36	0.65	0.40	**0.50**	0.33	0.71	**0.45**	0.79	0.34	0.47	0.34	0.64	**0.44**	0.24	0.66	**0.34**	0.435
$CBOW_{amazon}$	0.38	0.39	0.38	0.48	0.51	0.49	0.22	0.70	0.33	0.82	0.39	0.53	0.33	0.46	0.38	0.47	0.24	0.32	0.411
$SECO\text{-}PREC\text{-}NPMI_{wiki}$	0.41	0.44	**0.43**	0.46	0.23	0.30	0.34	0.44	0.38	0.66	0.64	**0.65**	0.48	0.39	0.43	0.27	0.49	**0.34**	0.422

ALM																
	ag-dg			*fr*			*hp*			*sd*			*sp*			Avg
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	
SG_{amazon}	0.58	0.44	**0.50**	0.40	0.55	0.46	0.77	0.31	0.44	0.40	0.71	0.51	0.23	0.33	0.27	0.440
$CBOW_{amazon}$	0.48	0.47	0.47	0.31	0.73	0.43	0.76	0.31	0.44	0.39	0.59	0.47	0.50	0.08	0.14	0.399
$SECO\text{-}PREC\text{-}NPMI_{wiki}$	0.52	0.27	0.36	0.42	0.52	**0.47**	0.64	0.70	**0.67**	0.57	0.53	**0.55**	0.24	0.33	**0.28**	0.464

ISEAR																
	ag			*dg*			*fr*			*hp*			*sd*			Avg
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	
SG_{amazon}	0.67	0.27	0.38	0.73	0.45	**0.56**	0.49	0.64	0.55	0.50	0.44	**0.47**	0.36	0.65	0.46	0.490
$CBOW_{amazon}$	0.42	0.62	0.50	0.61	0.47	0.53	0.53	0.67	**0.59**	0.60	0.32	0.42	0.51	0.47	**0.49**	0.510
$SECO\text{-}PREC\text{-}NPMI_{amazon}$	0.48	0.54	**0.51**	0.78	0.44	**0.56**	0.52	0.69	**0.59**	0.55	0.34	0.42	0.41	0.57	0.48	0.512

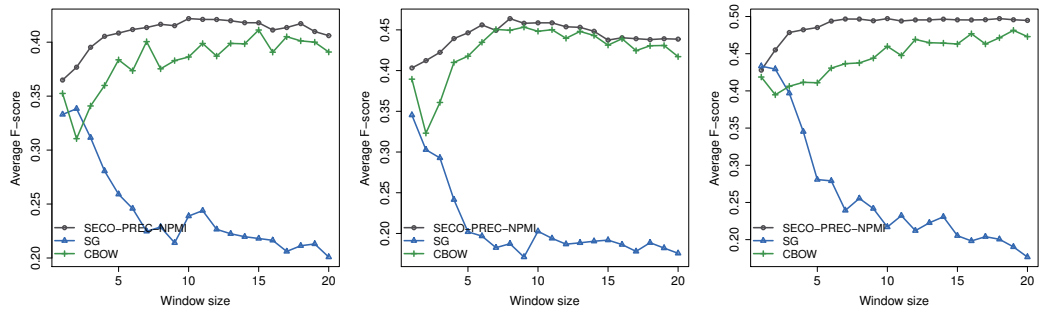
Table 5: Details of emotion category results for best window size/training corpus combination for three evaluation datasets. *ag* = anger, *dg* = disgust, *fr* = fear, *hp* = happy, *sd* = sad, *sp* = surprise. P = precision, R = recall, F = F-score.

iii) To analyze the individual emotion category results, we present the results of the best approach/training corpus combination in Table 5. In general, the *happiness* category obtains the highest results in two datasets while *fear* does best on the third. On the other hand, the most difficult category to be classified correctly seems to be *surprise*. One avenue of future work could include experimenting with various seed words to increase the accuracy of such emotions.

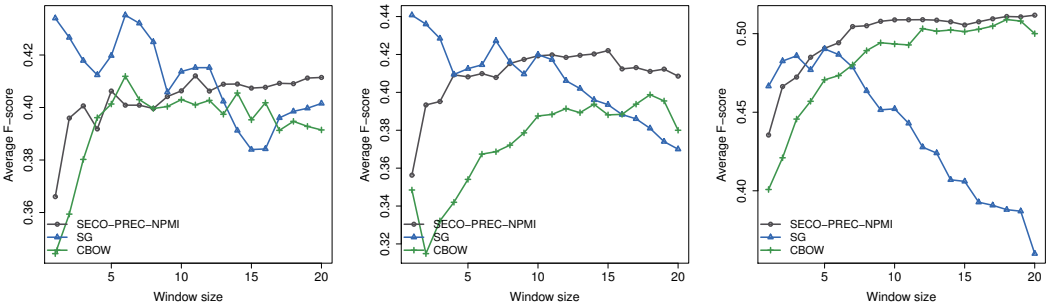
iv) Figure 3 summarizes the results of sensitivity of the three algorithms, SG, CBOW and `SECO-PREC-NPMI`, to different window parameter settings. On Wikipedia corpus, `SECO-PREC-NPMI` is consistently better than the others, whereas SG takes better advantage of the Amazon corpus. While `SECO-PREC-NPMI` and CBOW get better with bigger context windows, SG depicts the opposite trend, best on windows less than 5.

To summarize the results:

- Initialized using one seed word per emotion category, the measures of semantic relatedness yield better results than the emotion lexicons.
- When the window size is not known, in general, `SECO-PREC-NPMI` yields consistent promising results on all the evaluation datasets, while SG provides competitive results on one dataset.
- `SECO-PREC-NPMI` and CBOW yield better results with larger window sizes, whereas SG is best on windows less than 5.



(a) Aman (Wikipedia) (b) Alm (Wikipedia) (c) ISEAR (Wikipedia)



(d) Aman (Amazon) (e) Alm (Amazon) (f) ISEAR (Amazon)

Figure 3: Parameter sensitivity results for 20 window sizes

6 Conclusion

Emotion detection from text requires the degree of word-emotion association which is generally obtained from emotion lexicons or measures of semantic relatedness. While manual lexicons require considerable human time and effort, automatic techinques provide average performance. In this paper, we described a novel approach to automatically learning word-emotion association scores. Using just one seed word per emotion category, our proposed approach `SECO-PREC-NPMI` significantly outperformed three emotion lexicons and two state-of-the-art word embeddings models when trained using the Wikipedia text corpus.

As future work, we plan to further improve the accuracy of emotion classification by experimenting with a variety of seed words and also adapt `SECO` to other tasks that require association between two words.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable feedback. This research is funded in part by the Big Data Research, Analytics and Information Network (BRAIN) Alliance established by the Ontario Research Fund - Research Excellence Program (ORF-RE), Natual Sciences and Engineering Research Council of Canada (NSERC), and Social Sciences and Humanities Research Council of Canada (SSHRC).

References

Ameeta Agrawal and Aijun An. 2012. Unsupervised emotion detection from text using semantic and syntactic relations. In *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '12, pages 346–353, Washington, DC, USA. IEEE Computer Society.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-

based emotion prediction. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 579–586, Stroudsburg, PA, USA. ACL.

Ebba Cecilia Ovesdotter Alm. 2008. *Affect in Text and Speech*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

Jeanette Altarriba, Lisa M. Bauer, and Claudia Benvenuto. 1999. Concreteness, context availability, and imageability ratings and word associations for abstract, concrete, and emotion words. *Behavior Research Methods, Instruments, & Computers*, 31(4):578–602.

Saima Aman and Stan Szpakowicz. 2007. Identifying expressions of emotion in text. In *Proceedings of the 10th International Conference on Text, Speech and Dialogue*, TSD'07, pages 196–205, Berlin, Heidelberg. Springer-Verlag.

Doug Beeferman, Adam Berger, and John Lafferty. 1997. A model of lexical attraction and repulsion. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, ACL '98, pages 373–380, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40.

Bernard Brosseau-Villeneuve, Jian-Yun Nie, and Noriko Kando. 2010. Towards an optimal weighting of context words based on distance. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 107–115. Association for Computational Linguistics.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Werner De Bondt, Rosa M Mayoral, and Eleuterio Vallelado. 2013. Behavioral decision-making in finance: An overview and assessment of selected research. *Spanish Journal of Finance and Accounting/Revista Española de Financiación y Contabilidad*, 42(157):99–118.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Shichuan Du, Yong Tao, and Aleix M Martinez. 2014. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462.

Facebook. 2016. Reactions now available globally. Accessed: 2016-03-05.

Jianfeng Gao, Ming Zhou, Jian-Yun Nie, Hongzhao He, and Weijun Chen. 2002. Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 183–190, New York, NY, USA. ACM.

Zornitsa Kozareva, Borja Navarro, Sonia Vazquez, and Andres Montoyo. 2007. Ua-zbsa: A headline emotion classification through web information. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 334–337, Prague, Czech Republic, June. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2177–2185. Curran Associates, Inc.

GermÃ¡n Kruszewski Marco Baroni, Georgiana Dinu. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*, 1:238–247.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 43–52, New York, NY, USA. ACM.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

Saif M. Mohammad and Peter D. Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, CAAGET '10, pages 26–34, Stroudsburg, PA, USA. Association for Computational Linguistics.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. 29(3):436–465.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. 2007. Textual affect sensing for sociable and expressive online communication. In *Proceedings of the 2Nd International Conference on Affective Computing and Intelligent Interaction*, ACII, pages 218–229, Berlin, Heidelberg. Springer-Verlag.

Jessica Perrie, Aminul Islam, Evangelos Milios, and Vlado Keselj. 2013. Using google n-grams to expand word-emotion association lexicon. In *Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing - Volume 2*, CICLing'13, pages 137–148, Berlin, Heidelberg. Springer-Verlag.

Robert Plutchik. 2001. The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350.

Magnus Sahlgren. 2006. *The Word-space model*. Ph.D. thesis, University of Stockholm (Sweden).

LJ Shrum, Min Liu, Mark Nespoli, and Tina M Lowrey. 2013. Persuasion in the marketplace: How theories of persuasion apply to marketing and advertising. *The Sage Handbook of Persuasion: Developments in Theory and Practice*, pages 314–330.

Jacopo Staiano and Marco Guerini. 2014. Depechemood: a lexicon for emotion analysis from crowd-annotated news. *CoRR*, abs/1405.1605.

Carlo Strapparava and Rada Mihalcea. 2008. Learning to identify emotions in text. In *Proceedings of the 2008 ACM Symposium on Applied Computing*, SAC '08, pages 1556–1560, New York, NY, USA. ACM.

Carlo Strapparava and Alessandro Valitutti. 2004. WordNet-Affect: An affective extension of WordNet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1083–1086. ELRA.

Changhua Yang, Kevin Hsin-Yih Lin, and Hsin-Hsi Chen. 2007. Building emotion lexicon from weblog corpora. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 133–136. Association for Computational Linguistics.