Modeling the Impact of Modifiers on Emotional Statements

Valentina Sintsova[1], Margarita Bolívar Jiménez[2], and Pearl Pu[1]

[1] Ecole Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
valentina.sintsova@alumni.epfl.ch, pearl.pu@epfl.ch
[2] Universidad Politécnica de Madrid
Madrid, Spain
m.bolivar@alumnos.upm.es

Abstract. Humans use a variety of modifiers to enrich communications with one another. While this is a deliberate subtlety in our language, the presence of modifiers can cause problems for emotion analysis by machines. Our research objective is to understand and compare the influence of different modifiers on a wide range of emotion categories. We propose a novel data analysis method that not only quantifies how much emotional statements change under each modifier, but also models how emotions shift and how their confidence changes. This method is based on comparing the distributions of emotion labels for modified and non-modified occurrences of emotional terms within labeled data. We apply this analysis to study six types of modifiers (negation, intensification, conditionality, tense, interrogation, and modality) within a large corpus of tweets with emotional hashtags. Our study sheds light on how to model negation relations between given emotions, reveals the impact of previously under-studied modifiers, and suggests how to detect more precise emotional statements.

Keywords: Emotion Analysis, Text Mining, Modifiers, Twitter

1 Introduction

Emotions are present in our everyday actions and influence our decisions, behaviors, and relationships. For that reason, emotion identification is becoming increasingly important for developing marketing strategies [4], inferring user interests [15], and understanding personal well-being [27]. A common strategy for text-based emotion recognition is to learn the associations of lexical terms to the given emotions and to classify text based on their occurrences [2, 19, 32]. However, the true feeling expressed in the text can change under a variety of modifiers. Even the most explicit emotional terms, such as the word 'happy', can relate to another emotion when they occur in the scope of a modifier, such as in the phrase 'not happy'. Examples from Table 1 illustrate how different modifiers can lead to different effects on emotional statements. In order to detect emotions

Table 1. Illustrative examples of the effects of modifiers on emotional statements.

Example	Modifier	Effect
I'm **not** ***ashamed*** to say it	Negation	Shifts to another emotion
I feel **so** ***relieved*** now	Intensifier	Increases emotion intensity
I feel **a little** ***sad*** tonight	Diminisher	Decreases emotion intensity
I know I **should** be ***happy***	Modality	Eliminates the presence of emotion
I'll be ***sad*** **if** you leave	Conditionality	Refers to a non-experienced emotion
Do you ***love*** her**?**	Interrogation	Refers to a non-confirmed emotion
I **was** ***happy*** then	Past Tense	Refers to a non-present emotion

in text more correctly, we should be able to properly model the effects of such modifiers on emotions. Addressing this challenge is the subject of this paper.

Previous studies do not fully address how these common modifiers affect specific emotions. Most related works tend to treat the strongest modifiers, e.g. negation and intensification, only in terms of the change of polarity and intensity [8, 9, 13]. The effects of other modifiers are disregarded or blocked by removing the modified statements [29]. When models specify per-emotion effects, they are hand-coded [1], which makes their adaptation to other emotion categories or data from another domain more difficult.

This paper proposes a unified data-driven analytical framework for modeling the effects of different modifier types on fine-grained emotion categories that are defined as sets of associated emotional expressions. We quantify the impact of each modifier on each specific emotion using a novel data analysis method, which is based on investigating the distributions of emotion labels for modified and non-modified occurrences of emotional terms in social media. The source of our data is Twitter, from which we collect tweets having an emotional hashtag, viewed as the author's self-revealed emotion label.

This data-driven method derives the model of the modifiers' effects from the patterns of their usage in the large corpus of linguistic data that are automatically labeled with emotions. This makes our method *easily adaptable* to different emotion categories and modifiers, hence giving us a significant advantage over the hand-coding method.

Another contribution of our work lies in *detailed modeling* of the modifiers' effects. Our model not only quantifies the difference between emotions associated with modified and non-modified emotional statements, but also describes how each modifier shifts emotions and changes our confidence in the emotions' presence. In this way, our method produces a fine-grained emotion-based model of modifiers' effects, describing how each emotion changes under each modifier.

We applied this method to analyze the effects of six types of linguistic modifiers. We discovered that the effects of all these modifiers are emotion-specific, confirming that we need a detailed, per-emotion model when treating modifiers. Furthermore, our analysis demonstrated how emotions shift under negations and revealed that some largely ignored modifiers, such as modality and interrogation, can also shift emotions. Finally, we showed the potential of the proposed

modeling to find more precise emotional statements. All these findings lead to important implications for developing a modifier-aware emotion classification system.

In the remainder of this paper, we describe the modifiers that we study and the related works that model their effects. Then we present our quantitative framework for analyzing modifiers and introduce the used emotion model and lexicon, modifiers' detection methods, and Twitter data. The two result sections present the extracted effects of modifiers and analyze their modeling within emotion classification. Finally, we summarize our work and discuss future directions.

2 Studied Types of Modifiers and Related Work

This paper studies six linguistic modifiers that were previously discussed in the context of sentiment analysis and that were shown to affect the meaning of emotional statements at least for some of them.

Negation. Negation is the most studied modifier of sentiment polarity. In the simplest approach, researchers consider negation as a polarity reversal [24]. Several other studies have concluded that negation affects both polarity and intensity in the words that are within its scope [3, 8, 10, 13]. Other researchers found that negation's effects depend on the prior polarity of words and the used negation expressions [33]. In automatically learned systems for polarity and emotion classification, negation is treated by considering each negated term as a separate feature [12, 23]. Our model follows an under-studied emotion-based approach, where the effect of negation is modeled separately for each emotion category of modified terms. The antonym-based reversal of emotions under negation was shown to increase the accuracy of polarity classification [1]. However, the reversal of emotion is not as simple because of complex relations between emotional concepts and between their linguistic expressions. For example, the phrase "I don't love you anymore" implies rather *Sadness* than either original emotion of *Love* or its antonym emotion of *Hate*. We assume that negation of an emotional term may express the absence of any specific emotion, may refer to another emotion category, or may just change the intensity (or confidence) of the given emotion.

Intensification. Intensification terms change the intensity of emotional words. They form two classes: *intensifiers* that increase the intensity, such as 'very' or 'really' (also called *amplifiers*), and *diminishers* that decrease the intensity, such as 'less' or 'little' (also called *downtoners*). To treat intensification, some methods add (subtract) points from the valence score of sentiment terms if they are preceded by an intensifier (a diminisher) [11, 24]. Other methods associate each intensification term with a hand-coded multiplication coefficient representing its strength [28]. We hypothesize that neither intensifiers nor diminishers can change the original emotion category. However, we assume that the confidence in the emotion presence would change according to the direction of an intensity shift (an increase or decrease). For instance, in the sentence "I love you so much", *Love* emotion is intensified, and we can be more confident that it is present.

Modality. Modality is a linguistic construction used to distinguish non-factual situations (*irrealis* events) from situations that happened or are happening (*realis* events). Modal operators can express a degree of uncertainty or possibility, and can also be used to express desires and needs [1, 24]. Consider, for instance, the phrases "I will regret it" and "I should be angry with you". In these examples the presence of modal verbs conceals whether the writer actually experienced the referenced emotion or not. Some modal expressions can directly imply the absence of the referenced emotion, as in "I would have loved to see you". Most researchers that consider modality in sentiment analysis treat it as a polarity blocker, ignoring the occurrence of sentiment terms in its scope [24, 28]. Benamara et al. [3] show that modality affects the strength and the degree of certainty of the opinion words that are within its scope. Others suggest hand-coding coefficients for the change in certainty or confidence [21]. Liu et al. [16] further argue for including detected modality classes as separate classification features. In a manually crafted model of modality's effects on emotional expressions, researchers found that some modal verbs can even reverse emotions [1]. This suggests we need to further investigate the effects of modality on specific emotions.

Conditionality. Conditional sentences can also describe *irrealis* events, i.e. potential or hypothetical situations that are not yet known to happen. For instance, in the sentence "I'll be sad if you leave", the emotion *Sadness* is not yet experienced. Conditionality is rarely treated in sentiment analysis. One example work suggests that classifying sentiment in conditional sentences is challenging and argues for training a tailored classifier to deal with them [20]. To shed light on how to treat conditionality in emotion recognition, we will study its exact effects on fine-grained emotions.

Interrogation. Interrogation represents sentences where a question is asked. Similarly to conditionality, we cannot be certain whether the states or events mentioned in questions actually happened. Thus, interrogative sentences can change our confidence in detected emotion, as shown by the example question "Do you love me?" However, interrogation can also shift an original emotion to another one: e.g. the sentence "Are you mad at me?" implies rather *Worry* than *Anger*. As interrogative emotional statements are not common in the review texts, the effects of interrogation are traditionally neglected in sentiment analysis. One of the few exceptions is the work of Tabaoda et al. [28], who consider interrogation as a polarity blocker, along with modality and conditionality. Nevertheless, interrogation is frequent in personal communications and deserves further investigation.

Past Tense. Past tense describes situations that happened in the past. Therefore, we may be more certain that the stated emotion was experienced (a potential confidence increase). Yet, expressing emotion in the past may also mean that currently it is not experienced anymore (a potential for confidence decrease). These two phrases illustrate these effects: "I was happy with you" and "I loved

you so much before". Thus, we investigate the effects of past tense to identify which case is more frequent and to conclude whether this under-studied modifier is relevant for consideration in emotion recognition.

Summary. Our analysis is different from the previous ones in the following aspects. First, we study six different linguistic modifiers using the same analytical framework, thus giving an advantage to compare their relative impact. Second, we consider their effects using a fine-grained emotion model of up to 20 categories. This choice helps us to model more precisely the emotional shifts of expressions that employ modifiers. Third, our modeling technique is data-driven and automatic, hence overcoming the costly nature of manual approaches. It also enables researchers to easily adapt, validate, and extend our analysis.

3 Quantitative Analysis of Modifiers

This section describes the method we have developed to quantify and analyze the effects of modifiers on emotional expressions.[3] We introduce below the model of emotion categories and the corresponding lexicon of explicit emotional terms that we use in our study. We describe the collected data that are automatically labeled with those emotions and present our approach to detect modifiers for the input emotional terms. Finally, we explain how these components are employed altogether to quantify the impact of different modifiers' types on the original emotion expressed by an emotional term.

3.1 Input Data

Emotion Model. In order to analyze a wide range of emotions, we chose the fine-grained emotion model of the Geneva Emotion Wheel (GEW, version 2.0 [25, 26]), which has twenty categories of emotions. The high number of categories enables a detailed modeling of the modifiers' effects, where we are likely to detect which emotions correspond to statements with modifiers. GEW was developed by the psychologists in order to categorize self-reported emotional states. It contains 10 positive and 10 negative emotion categories, each one represented by two close category names (e.g. Amusement/Laughter). We will use only the first names throughout the paper.

Lexicon of Explicit Emotional Terms. A list of explicit emotional terms is associated with this model—the Geneva Affect Label Coder (GALC) [25]. It is an affective lexicon that enumerates for each emotion category the stemmed words expressing it, i.e. each stemmed word from the GALC lexicon is associated with an emotion category. Overall, there are 212 stems for 20 GEW emotions, 10.9 in average per each emotion category. However, we discovered that using

[3] The input data, used linguistic resources, and code for analysis method are available at www.cicling.org/2017/data/263/.

stems with a wildcard token * at the end is undesirable, as sometimes non-related terms would be also matched. For instance, one of the most frequently mapped instances of the GALC stemmed term *happ** (*Happiness* emotion) is *happy*, which is the correct association, but the instance *happen* is also frequent while it does not correspond to this emotion category. This is why we instantiate the stemmed words into actual linguistic tokens by matching those stems in the dataset of around 15 million random tweets collected with the Twitter Sample API in Nov. and Dec. 2014. Then, we manually discovered correctly matched emotional terms among the most frequent instances. The new revised lexicon GALC-R consists of 1026 terms, 52.9 in average per emotion category.

Twitter Data Labeled by Emotional Hashtags. To perform our data-driven quantitative analysis, we require a large dataset. Furthermore, this dataset must be labeled with emotions. As manual annotation is not achievable at the desired scale, we resort to using the pseudo-annotated dataset of tweets. To obtain such dataset labeled with GEW emotion categories, we follow the distant supervision idea of using the emotional hashtags appearing at the end of the text as a self-reported emotion label for the tweet [7, 18, 31]. Concerning the quality, we rely on the previous evaluations of similar hashtag-based labeling, which showed that the emotion of the hashtag correctly corresponded to the tweet content in 83% of tweets for a large set of emotional and mood-descriptive hashtags [6].

We specify the list of 167 emotional hashtags assigned to the GEW categories based on previously introduced GALC lexicon [25]. 17.6 millions of English tweets with those hashtags were collected via Twitter Streaming API in Mar.–May of 2014. After cleaning, we extracted 1, 729, 980 tweets that had those hashtags at the end of the text, were not repeated, were not retweets, did not contain URLs, and were assigned to only one emotion category. All these tweets were converted to lower-case and preprocessed to correctly separate emoticons, usernames, and punctuation marks from other tokens. We randomly sampled 1.5 million of such tweets to be used for studying the effects of the modifiers on emotional terms (analysis dataset D_A). The remaining 229, 980 tweets will be used in the emotion classification experiments (test dataset D_T).

3.2 Data Preparation

Our data preparation process consists of three main steps as shown in Figure 1. We overview each of them and present our terminology.

First, we take as input the analysis dataset D_A, which consists of the ***tweets with emotional hashtags***. For each hashtagged tweet, we consider the emotion category associated with the emotional hashtag to be a true emotion label for the full tweet (one per tweet), and refer to it as a *hashtag emotion*.

We define the *emotion distribution* of a subset of the hashtagged tweets as the distribution of the tweets' *hashtag emotion* labels over the GEW's 20 categories. For each category, we compute the proportional amount of tweets

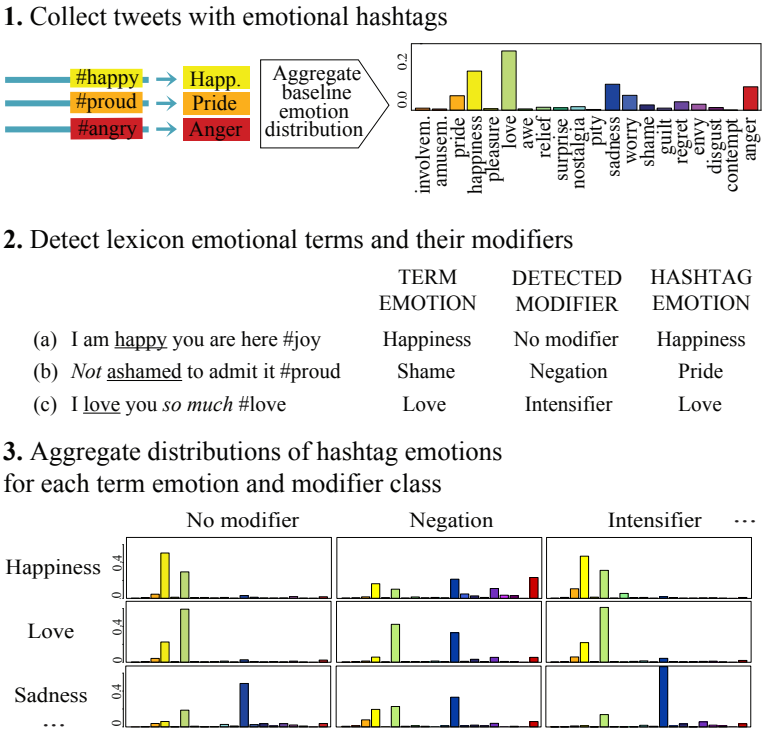

Fig. 1. The process of data collection, preparation, and extraction of emotion distributions.

whose emotional hashtags belong to that category. The emotion distribution of all tweets in the analysis dataset D_A is called the *baseline distribution* P_{BASE}.

Next, we ***detect lexical emotional terms and their modifiers*** in the text of the input tweets. We identify a subset of the above dataset with tweets containing exactly one emotional term from the emotion lexicon GALC-R. We look for such terms in the content of the tweets while disregarding their emotional hashtags. There are $245,591$ such tweets, some of them containing modified emotional terms and some non-modified ones.

For each tweet in this set, we construct a triplet data representation with the following elements: 1) the GEW category of the detected emotional term (henceforth called *term emotion*), 2) whether that term is modified and with which modifier (*non-overlapping modifier class*), and 3) the true emotion as revealed by the tweet's hashtag (*hashtag emotion*). For example, from the tweet "I don't love it #sad" we will identify *Love* as its *term emotion* (based on the term *love*), Negation as the *modifier class* (based on the negation term *don't*), and *Sadness* as the *hashtag emotion* label (based on hashtag *#sad*).

Finally, we ***compute the emotion distributions for each term emotion and each modifier class***. For every *term emotion* such as *Happiness*, we con-

struct emotion distributions of 20 *hashtag emotions* for every *non-overlapping modifier class*, starting with a distribution for tweets without any modifiers, for tweets with negations, for tweets with intensifiers, etc. Note that for each *term emotion* we aggregate all the tweets with the lexicon terms corresponding to that emotion.

We further provide details about detecting modifiers, separating modifier classes, and extracting emotion distributions.

Detection of Modifiers that Affect Lexicon Terms. We apply the modifier detection module to discover how emotional terms are modified by the respective modifier types. We detect each of six modifiers' types based on the presence of specific words and multi-word expressions from a modifier's list. Depending on the modifier's scope, these terms can appear either some words before or after an emotional term in question. We additionally ensure that no punctuation marks or emoticons appear in between the modifier and emotion terms to avoid splitting sentences. Also, to avoid detection errors, we compile lists of frequent false positive expressions and ignore modifier terms that appear within them.

Negation. The list of negation expressions contains common negation words, such as *wasn't*, *not*, or *no*, and their misspelled variants, such as *werent* or *didnt* (taken from [1]). It also includes 38 verbs, such as *pretend* or *fail*, implying that a modified statement is not experienced or does not happen (taken from [17], where they are marked as having a negative signature). To be detected, a negation word should appear up to three words before the emotional term. We additionally extract 202 false positive expressions for negation, such as *nothing but* and *can't help*, among which 83 are marked as intensifiers, e.g. *couldn't be more*. We also deal with double negations, which are marked as not negated.

Intensification. We compile the lists of 93 intensifiers (e.g. *much*) and 38 diminishers (e.g. *a bit*) from the related literature [1, 9], and extend them with manually validated frequent n-grams containing those words. We further classify each term according to its position: whether it can appear before (e.g. *lots of*) or after (e.g. *very much*) the emotional term, or both (e.g. *less*). The scope of an intensification modifier is then defined as one word directly before or after it, depending on its classification. 9 false positive phrases, such as *that kind of* and *at least*, were also added.

Modality. Our list of modal expressions has 143 terms. The significant part of it consists of modal verbs, such as *should*, *might*, or *can*. *Will*, *'ll*, *wont* are also in this list, i.e. the emotional terms in the future tense will be detected as modified with modality [22]. Additionally, our list contains the expressions of desire (e.g. *wish*, *want*) and of uncertainty (e.g. *maybe*, *seems*). To be detected, a modal expression should appear up to 4 words before an emotional term. Note that we avoided including modal expressions of high certainty or 'trueness', such as *sure* or *indeed*, because they are assumed to have a different effect on emotions than other considered modal expressions [21].

Interrogation, Conditionality, and Past Tense. We detect interrogation by inspecting whether there is the interrogation sign '*?*' after the emotional term or the question-specific patterns, such as *am i* and *why does*, before the term. Conditionality is detected by finding the word '*if*' before the emotional term. The sentence boundaries are checked in both cases. To detect past tense, we search for the part-of-speech tag specific for verbs in the past tense, using Stanford POS Tagger [30]. The emotional term is considered to be in past tense if this tag appears up to four tokens before it.

Separation of Modifier Classes. Several modifiers can modify the same term in the text, e.g. both negation and intensification are present in the phrase "not very interested". To exclude confounding effects between modifiers from our analysis, we split the entries of modified terms into *non-overlapping modifier classes* using the following rules. We recognize Past tense modifier only if it does not overlap with any other modifier, otherwise we assign the overlapping modifier alone (i.e. Past Tense plus Negation will be assigned to the Negation class). The case of Modality and Conditionality is assigned to Conditionality only. The same is for Modality and Interrogation (assigned to Interrogation only). We also separate a class of Mixed Negation containing all the cases where other modifiers (except for Past Tense) overlap with Negation. All other overlapping cases of found modifiers are placed into the Mixed class and are not considered in the analysis of the modifiers' effects.

We note that 34% of the emotional terms from GALC-R are modified by at least one modifier, with Intensifiers being the most common modifier (14.9% of the entries), followed by Past Tense (5.2%). Negations in total modify 3.6% of the terms, while 32% of them are Mixed Negation cases. Mixed class covers only 2.2% of the terms.

Extracted Emotion Distributions. Our data preparation step produces three types of *emotion distributions* (where each of the emotion distributions represents the proportions of hashtag emotions in the corresponding subsets of tweets):

1) The *baseline distribution* P_{BASE}, which is the emotion distribution of the entire dataset D_A.

2) The *modified emotion distribution* $P_M(E)$, which is the emotion distribution of the tweets with the term emotion E and with the non-overlapping modifier class M. We count only those tweets where an emotional term for emotion E is within the scope of the modifier M.

3) The *non-modified emotion distribution* $P(E)$, which is the emotion distribution of the tweets with the term emotion E and having no modifiers. In order to neutralize potential mistakes of the modifiers' scope detection process, we exclude from this distribution the tweets that contain any modifiers' terms, ignoring whether emotional terms are within their scope or not.

Our analysis of the effects of each modifier class will be based on comparing these emotion distributions.

3.3 Quantification of Modifiers' Impact

For each modifier class, we study how it affects each *term emotion* by estimating the change in the emotion distributions of the tweets with modified and non-modified terms. We quantify the influence of the modifiers by comparing the corresponding distributions of hashtag emotion labels using the Kullback-Leibler (KL) divergence [14, 5].

The KL divergence is an asymmetrical measure of the difference between two probability distributions S and Q. In our discrete case, it is computed as follows:

$$D(S||Q) = \sum_i s_i \ \log \frac{s_i}{q_i}$$

where s_i and q_i are the corresponding percentage of hashtag emotion E_i in the emotion distributions S and Q. The KL divergence measures how well the distribution S could be approximated by the distribution Q. The closer it is to zero, the better is the approximation. As our goal is to analyze the modified distributions, we consider more restrictive modified distributions $P_M(E)$ as S in the formula, and take more general non-modified or baseline distributions as Q.

To obtain representative *modified emotion distributions*, we include in the analysis of a modifier only those emotions for which at least 50 tweets contain their modified terms. Also, to avoid division by zero in the KL computation when emotion distributions are sparse, we add a smoothing constant of 0.05 to each emotion label count before normalizing the distributions to percentage values.

Our analysis of modifiers' effects aims to answer the following three questions regarding the effects of each modifier class M on a specific term emotion E, which we will refer to as original term emotion.

Question 1. To what extent does the modified emotion differ from the original non-modified emotion? **(modifier divergence)**

We answer this question by comparing the emotion distributions of the modified cases with the ones without any modifier (i.e. modified distribution vs. non-modified distribution for the original term emotion E). This means we compute the KL divergence $D\big(P_M(E)||P(E)\big)$. We refer to this metric as a *modifier divergence*. It can help quantify how much impact each modifier has. In this comparison, we assume that people who express their emotions with and without using modifiers assign emotion labels to their statements in a similar manner.

Question 2. Does the original emotion change under the modifier into another outcome emotion, or does it stay the same? **(shift** or **no shift)**

To detect which non-modified emotion approximates the best the extracted modified emotion, we compare the distribution of the modified emotion $P_M(E)$ with distributions of each non-modified emotion $P(E_i)$. The emotion E_i that provides the minimal KL divergence will be referred to as the *outcome emotion* E_{out} under that modifier, i.e.

$$E_{out} = \arg\min_{E_i} D\big(P_M\left(E\right)||P\left(E_i\right)\big).$$

We say that the modifier *shifts* the emotion E if the outcome emotion is different from the original emotion, i.e. if $E_{out} \neq E$. Otherwise, we say the emotion remains the same or *no shift* has been detected under the modifier ($E_{out} = E$). This knowledge is necessary for properly modeling the modifier's effect within emotion classification.

Question 3. How confident are we that the discovered outcome emotion is actually expressed in the modified text? **(confidence coefficient)**

Regardless of whether there was a shift of emotion or not, it is likely that the modified distribution $P_M(E)$ differs from the closest non-modified distribution of the outcome emotion $P(E_{out})$. It can differ in two ways: the modified emotion distribution can be more pronounced than the non-modified one, e.g. due to a higher peak for the outcome emotion; or it can have a more random distribution, corresponding to a more mixed state of emotions or an absence of them. The first case intuitively increases our confidence that the outcome emotion is present, while the second one decreases it.

Following this intuition, we compute a *confidence coefficient* (CC) that measures a change of confidence in the presence of the outcome emotion in modified distribution relative to such confidence in the non-modified case. To compute it, we additionally compare both modified and non-modified distributions with the baseline emotion distribution of all analysis tweets P_{BASE}. We define the *confidence coefficient* (CC) as a ratio of two KL divergences: one between the modified and baseline distributions and one between non-modified distribution of the outcome emotion E_{out} and the baseline distribution, i.e.

$$CC = \frac{D\big(P_M(E)||P_{BASE}\big)}{D\big(P(E_{out})||P_{BASE}\big)}$$

The confidence decrease ($CC < 1$) implies that the modified emotion distribution $P_M(E)$ is more random than the non-modified $P(E_{out})$. And the confidence increase ($CC > 1$) implies that the modified emotion distribution $P_M(E)$ is more pronounced than the non-modified one.

To illustrate the suggested analysis method, we visualize the corresponding emotion distributions in the case of analyzing how Negation modifier affects original emotion of *Pride* (Figure 2). It can be observed that the distribution for negated Pride (B) is considerably different from the non-modified Pride distribution (A). More particularly, it has the peak on *Shame* instead of *Pride*, which leads to a high modifier divergence value of 1.96. Furthermore, this makes the non-modified Shame distribution (C) to have the smallest KL divergence to the negated Pride distribution. We thus infer that *Shame* is the outcome emotion of *Pride* under Negation. However, negated Pride distribution (B) has higher percentage of *Pride* and *Love* than non-modified Shame distribution (C), showing that it does not follow it exactly. In result, (B) is closer to the baseline distribution P_{BASE} than (C) (1.02 vs. 1.11), which in its turn results in decreased confidence ($CC = 0.92$).

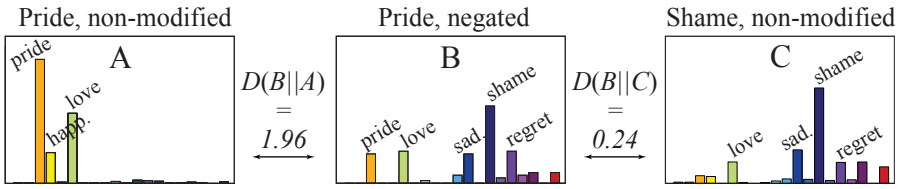

Fig. 2. Examples of non-modified (A) and modified (B) emotion distributions for analyzing the effects of Negation on Pride. (C) visualizes the non-modified distribution of the outcome emotion Shame, which has the smallest KL divergence to (B).

3.4 General Remarks

The presented method of analysis does not aim at building a general theory of modifiers' effects on emotions. Instead, it provides a data-driven linguistic approach where emotion categories are detected based on corresponding sets of emotional terms or expressions. Because of that, the extracted model of modifiers' effects is purely linguistic and the exact computed effects depend on the used set of emotional expressions, applied methods of modifiers' detection, and input linguistic data. The goal of this analysis is to better understand and model how modifiers affect emotional statements and how such effects could be properly treated within emotion classification.

4 Computed Effects of Modifiers

Our method models the effects of each modifier class M on each emotion E in terms of four characteristics: modifier divergence, the outcome emotion, whether there is an emotion shift, and the confidence coefficient of the outcome emotion. In this section, we summarize the detected effects of modifiers.

To show how each modifier affects the explicit emotional statements in general, we present the aggregated effects of modifiers in Table 2. We report for each modifier the average of *modifier divergences* across all emotion categories along with the names of the original and outcome emotions corresponding to the highest modifier divergence. We also summarize the behavior of shifts and confidence changes: what proportion of the original emotions shifts into other outcome emotions with either increase or decrease of confidence, and what proportion of emotions remains the same under the modifier. Note that we use in our analysis (and thus in this aggregation) only the emotion categories for which enough modified entries are detected (≥ 50).

These results confirm the expected differences in the impact of different modifiers, with Intensifiers being the least influential modifiers and Negation—the most. At the same time, they show that the effects of each modifier differ depending on the emotion category it modifies. This is reflected in the facts that every modifier shifts at least one emotion and that no modifiers have the same effect on all emotions. According to the overall shifting pattern, we can separate three groups of effects: no shift, mixed, and shift.

Table 2. Comparison of the different non-overlapping modifier classes using metrics aggregated across emotion categories. The modifier divergence (MD) for an emotion category is the KL divergence between modified and non-modified distributions. We count the percentage of shifted and non-shifted emotions under each modifier, aggregated by the confidence coefficient (CC) behavior.

Modifier class	MD mean	$E \to E_{out}$ for max MD	% of shifted		% of no-shift		Summary of effects
			$CC > 1$	$CC < 1$	$CC > 1$	$CC < 1$	
Intensifiers	0.12	Nostal. → Regr.	0%	11%	**78%**	11%	no shift, $CC > 1$
Past Tense	0.17	Guilt → Guilt	0%	6%	19%	**75%**	no shift, $CC < 1$
Modality	0.19	Involv. → Worry	6%	13%	6%	**75%**	no shift, $CC < 1$
Conditionality	0.26	Involv. → Sadn.	0%	27%	18%	**55%**	no shift, $CC < 1$
Diminishers	0.28	Nostal. → Regr.	11%	11%	**56%**	22%	no shift, $CC > 1$
Interrogation	0.40	Awe → Involv.	29%	24%	**35%**	12%	mixed
Mix. Negation	0.52	Pleas. → Regr.	8%	**50%**	0%	42%	mixed
Negation	0.80	Pride → Shame	0%	**75%**	0%	25%	shift, $CC < 1$

Modifiers with No-Shift Effects. The smallest value of average modifier divergence belongs to Intensifiers. As expected, for most emotions they do not shift the original emotion, but increase its confidence ($CC > 1$).

Past Tense, Modality, and Conditionality modifiers have another behavior. They mostly decrease confidence ($CC < 1$) without shifting the original emotion. This means that these modifiers can introduce uncertainty on whether a specific emotion is expressed.

Our model further makes an interesting but counter-intuitive observation about Diminishers: they increase confidence for most of emotions, while preserving the original emotion. This is explained by the fact that when a person states he or she is "kinda/a little/only/a bit" "sad/in love/worry/disappointed" we can be more confident that the stated emotion is actually being experienced.

Yet, there are exceptions from these main patterns of behavior. For example, some negative emotions expressed in the past tense are linked more confidently to the associated emotions, i.e. with $CC > 1$ (an example is "I was disappointed"). Also, each of these modifiers with general no-shift effects shifts some emotions. For example, Modality shifts 19% of emotions (3 out of 16 analyzed ones) and Conditionality—27% (3 out of 11). We note that some of these shifts reflect the specifics of using the studied emotional expressions in English. For instance, Modality shifts *Involvement* into *Worry/Fear* with an increased confidence due to the widespread of the phrase "this will/shall/gonna/should be interesting" that expresses the author's worry about what is going to happen.

Modifiers with Mixed Effects. Not every modifier has a clear overall shifting behavior. More particularly, we discover that Interrogation has its own pattern. It shifts many positive emotions into *Involvement*. This may be explained by the fact that asking other people questions about their positive emotions is

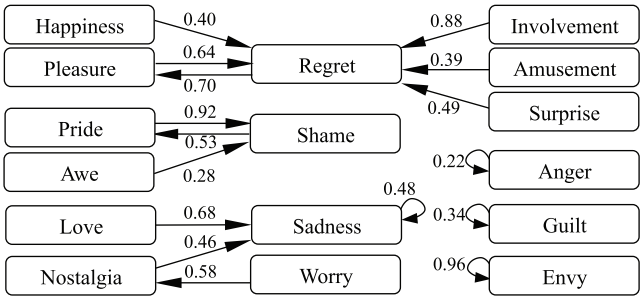

Fig. 3. The extracted model of emotion shifts under negation. The arrows point to the outcome emotion of negating the original emotion. They are labeled with the confidence coefficients (CCs) of the outcome emotions.

an expression of *Involvement/Interest* by itself. Meanwhile, negative emotions mostly do not shift in interrogative sentences.

Mixed Negation has a mixed effect as well. For example, it shifts several positive emotions, including *Happiness* and *Pleasure*, into *Regret*, because of the dominance of Negation mixed with Modality (an example is "I can't be happy"). At the same time, many other emotions stay the same with the lower confidence.

Modifiers with Shifting Effects. The highest modifier divergence value corresponds, as expected, to Negation. In line with the previous findings in sentiment analysis, we observe that negation tends to shift emotions (it happens for 12 out of 16 analyzed emotions, i.e. for 75%) and decreases the confidence in the outcome for all emotions, even non-shifted (average CC is $0.56 < 1$). However, our analytical method allows establishing which emotion the original emotion shifts to (i.e. the outcome emotion), and discovering emotions that do not shift even after being modified. Figure 3 summarizes such per-emotion shifting effects of negation. We can observe several clusters of these effects.

1) Five of the positive emotions shift towards *Regret*, while *Regret* itself shifts back towards *Pleasure*. This cluster represents a standard notion of negation influence, where "not happy" and "not amused" are considered to have negative sentiment. It is noteworthy that *Happiness* does not shift to its direct antonym *Sadness*. Also, we do not have direct antonyms of *Amusement* and *Involvement* in the emotion model, thus under negation they shift into the most appropriate emotion category among the given ones (i.e. *Regret* in this case).

2) We discover a reciprocal negation relationship along the antonym pair *Pride-Shame*. *Awe* also shifts towards *Shame*, which can be attributed to the frequently negated expression "no wonder".

3) Negation of *Love* and *Nostalgia* becomes *Sadness*, as in the tweet "Nobody loves me enough to hang out with me". At the same time, *Worry* shifts into *Nostalgia*. However, the KL divergence between modified and baseline emotion

distributions is small and thus negated *Worry* might rather represent a mixture of emotions than *Nostalgia*, even with a lower confidence.

4) There are four negative emotions, namely *Sadness*, *Anger*, *Envy*, and *Guilt*, for which there is no emotion shift under negation (i.e. they remain the same). This can be illustrated by the examples "I'm not normally an angry ranty person" and "I'm trying not to get sad". The confidence coefficients are small for all of these emotions, except *Envy*, for which it is close to one, meaning that "not envious" has almost the same meaning as "envious".

Overall, all positive emotions shift towards negative emotions, and several negative emotions shift towards positive ones. This confirms the expected power of negation to reverse polarity of emotions. Yet, we find no shift under negation for several negative emotions. This once again shows the importance of treating the effects of modifiers individually for each emotion.

5 Classification Quality of Modified Statements

We evaluate in this section how the classification quality of emotional statements depends on the presence of modifiers and the type of their modeled effects.

The extracted quantitative model of the modifiers' effects specifies, for each emotion category and a modifier, what is the outcome emotion after modification and what is its confidence coefficient. For example, it specifies that a negated term of *Sadness* remains assigned to the category *Sadness* with a confidence coefficient of 0.48.

As the basis of classification, we use the GALC−R lexicon of explicit emotional terms. For each occurrence of a lexicon term, we detect which of the studied modifiers are present. We again use the non-overlapping classes of detected modifiers (as described in Section 3.2). Based on the presence of modifiers and their extracted effects, we separate three cases of emotional terms' occurrences:

1) ***Not Modified***—No modifiers are detected. We return the original emotion associated with the emotional term.

2) ***No Shift***—Exactly one modifier is detected for the term, and it produces no shift of the emotion associated with the term. The term's emotion is returned.

3) ***Shift***—Exactly one modifier is detected for the term, and it shifts the original term emotion into another outcome emotion. We separate two scenarios for treating this shift: whether to return the outcome emotion or the original emotion of the emotional term.

We exclude the mixed cases, where several non-overlapping modifiers are detected, and the cases where the modifier's effect is not modeled because not enough of such modified statements appeared in the analysis dataset D_A.

To compute the classification quality of each case of modified emotional statements, we use a test dataset of hashtagged tweets D_T, containing $229,980$ tweets with one of the emotional hashtags for 20 GEW emotion categories. These tweets did not participate in the extraction of the modifiers' effects. We again consider the emotion category of a hashtag to be a ground-truth label, and remove the

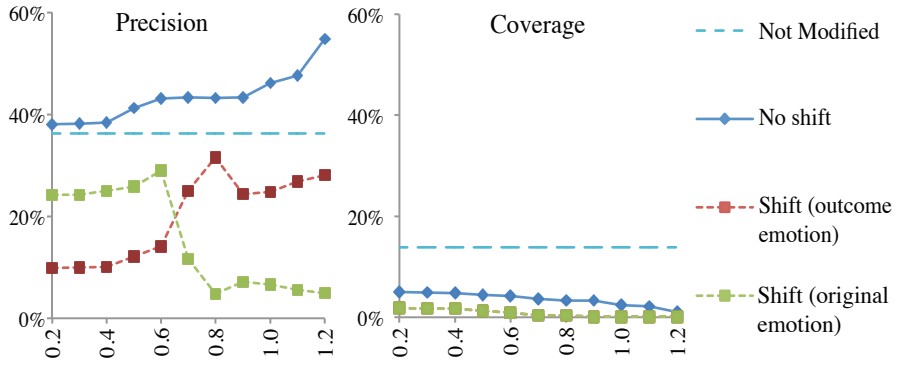

Fig. 4. The precision and coverage of different modified cases of emotional terms depending on the confidence threshold τ. Only emotional terms with $CC \geq \tau$ are included in each case.

hashtags themselves from the tweet text. We use precision and coverage as performance metrics. Precision is the ratio of correctly found hashtag labels among all labels returned based on the considered statements. Coverage is the percentage of tweets in which the considered statements are found.

We investigate the quality of classification depending on the level of filtering: we ignore the modified statements with the confidence coefficient CC lower than a confidence threshold τ. Figure 4 shows the dependency of precision and coverage on the confidence threshold τ for the three considered modified cases. When $\tau = 0.2$ all corresponding modified statements are used without filtering. Notice that the non-modified case is independent of τ values.

The results show that no-shift modified cases have a higher precision than non-modified emotional entries for any values of τ. This means that when an emotional term appears in the scope of a no-shift modifier the precision of its association with the corresponding emotion is higher. Also, we can observe that the higher the confidence threshold τ is, the higher the precision of the no-shift modified cases is, but the lower their coverage is. We can thus identify more precise emotional statements by increasing the τ value.

Considering the shifted modified cases, we observe that their precision is lower than of the non-modified case, regardless of what emotion (original or outcome) is returned. This means that we can exclude such shifted cases altogether in order to obtain more precise classification results. The plot also shows a shift in dominance between two options to return emotion at $\tau = 0.7$. This suggests how to potentially increase the overall precision without excluding shifted modified cases: we can return the original emotion for lower CC values, and the shifted, outcome emotion for higher CC values.

In essence, knowledge about the shifting and non-shifting behavior of modified cases helps us find more precise emotional statements. Therefore, we can construct higher-precision classifiers, which can be then used to initialize distant

supervision algorithms or to identify more reliable classification examples within an application.

Limitations In the current analysis, we considered only lexicon-based classification approach and only one lexicon (GALC-R) for which the modifiers' model was computed. Further research is required to understand how to incorporate such modifiers' model within machine learning-based classification methods, such as Support Vector Machines or Multinomial Naïve Bayes, and how this approach relates to other automatic techniques of treating modifiers, such as coding them as separate features.

6 Conclusion and Future Work

This paper proposes a data analysis method to model the effects of different linguistic modifiers on fine-grained emotional statements based on their usage in social media. It analyses how the modifiers change respective emotion distributions, how they shift the term emotions, and how they affect the confidence in the outcome emotions. With this method, we study the effects of six different linguistic modifiers, such as negation, intensification, and modality, on the explicit emotional terms that are within their scope. As labeled data, we use a large number of tweets with author's self-revealed emotions, identified via emotional hashtags. This work, to our best knowledge, is the first systematic study of the effects of different modifiers at a fine-grained level of emotion categories.

Our analysis reveals multiple interesting patterns of modifiers' impact. First of all, the effects of modifiers are non-uniform across emotion categories, suggesting that to more effectively treat modifiers and their effects we need to model the fine-grained per-emotion effects. For example, we show that some under-studied modifiers can even shift emotion categories: conditionality and modality shift *Involvement* to *Sadness* and *Worry* correspondingly. Second, our data confirms that negations are the most notorious modifiers, shifting 75% of the emotion categories. More interestingly, our model shows how the original emotions shift in the presence of negation and other modifiers. Third, we show the potential of incorporating the computed modifier model along with its confidence coefficients to identify more precise emotional statements. Such profound, detailed understanding of the modifiers' effects is essential for building emotion classifiers of superior quality.

The proposed method aims at helping researchers to treat modifiers for classification purposes, not at universal modeling of emotion-modifiers relations. Nevertheless, our data-driven modeling method can extract the different modifiers' effects within any data where bootstrapping a large quantity of high-quality emotional data is feasible, e.g. using hashtags or emoticons. It also allows updating the modifier model for new modifier types or emotion categories, which would help to test another hypothesized impact. Because of these properties, our analytical framework could facilitate future research on automatic discovery of new modifier expressions, investigation of other linguistic or contextual modifiers, and construction of modifier-aware emotion classification systems.

References

1. Carrillo-de Albornoz, J., Plaza, L.: An emotion-based model of negation, intensifiers, and modality for polarity and intensity classification. Journal of the American Society for Information Science and Technology 64(8), 1618–1633 (2013)
2. Aman, S., Szpakowicz, S.: Identifying expressions of emotion in text. In: Proceedings of Text, Speech and Dialogue Conference (TSD), Volume 4629 of LNCS. pp. 196–205. Springer (2007)
3. Benamara, F., Chardon, B., Mathieu, Y., Popescu, V., Asher, N.: How do negation and modality impact on opinions? In: Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics. pp. 10–18. ACL (2012)
4. Consoli, D.: A new concept of marketing: The emotional marketing. BRAND. Broad Research in Accounting, Negotiation, and Distribution 1(1), 52–59 (2010)
5. Cover, T.M., Thomas, J.A.: Elements of information theory. John Wiley & Sons (2006)
6. De Choudhury, M., Counts, S., Gamon, M.: Not all moods are created equal! Exploring human emotional states in social media. In: Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM). pp. 66–73 (2012)
7. De Choudhury, M., Gamon, M., Counts, S.: Happy, nervous or surprised? Classification of human affective states in social media. In: Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM). pp. 435–438 (2012)
8. Hogenboom, A., Van Iterson, P., Heerschop, B., Frasincar, F., Kaymak, U.: Determining negation scope and strength in sentiment analysis. In: Proc. of Systems, Man, and Cybernetics (SMC). pp. 2589–2594. IEEE (2011)
9. Hutto, C.J., Gilbert, E.: VADER: A parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM). pp. 216–225 (2014)
10. Jia, L., Yu, C., Meng, W.: The effect of negation on sentiment analysis and retrieval effectiveness. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM). pp. 1827–1830. ACM (2009)
11. Kennedy, A., Inkpen, D.: Sentiment classification of movie reviews using contextual valence shifters. Computational intelligence 22(2), 110–125 (2006)
12. Kiritchenko, S., Zhu, X., Mohammad, S.M.: Sentiment analysis of short informal texts. Journal of Artificial Intelligence Research pp. 723–762 (2014)
13. Kou, X.: The effect of modifiers for sentiment analysis. In: Chinese Lexical Semantics, pp. 240–250. Springer (2014)
14. Kullback, S., Leibler, R.A.: On information and sufficiency. The annals of mathematical statistics 22(1), 79–86 (1951)
15. Lewenberg, Y., Bachrach, Y., Volkova, S.: Using emotions to predict user interest areas in online social networks. In: Proceedings of International Conference on Data Science and Advanced Analytics (DSAA). pp. 1–10. IEEE (2015)
16. Liu, Y., Yu, X., Chen, Z., Liu, B.: Sentiment analysis of sentences with modalities. In: Proceedings of UnstructureNLP. pp. 39–44. ACM (2013)
17. Lotan, A., Stern, A., Dagan, I.: TruthTeller: Annotating predicate truth. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). pp. 752–757. ACL (2013)
18. Mohammad, S.M.: #Emotional tweets. In: Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM). pp. 246–255. ACL (2012)

19. Mohammad, S.M., Turney, P.D.: Crowdsourcing a word–emotion association lexicon. Computational Intelligence 29(3), 436–465 (2013)
20. Narayanan, R., Liu, B., Choudhary, A.: Sentiment analysis of conditional sentences. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. vol. 1, pp. 180–189. ACL (2009)
21. Neviarouskaya, A., Prendinger, H., Ishizuka, M.: SentiFul: A lexicon for sentiment analysis. IEEE Transactions on Affective Computing 2(1), 22–36 (2011)
22. Palmer, F.R.: Modality and the English modals. Routledge (2014)
23. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: Sentiment classification using machine learning techniques. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. vol. 10, pp. 79–86. ACL (2002)
24. Polanyi, L., Zaenen, A.: Contextual valence shifters. In: Computing attitude and affect in text: Theory and applications, pp. 1–10. Springer (2006)
25. Scherer, K.R.: What are emotions? And how can they be measured? Social science information 44(4), 695–729 (2005)
26. Scherer, K.R., Shuman, V., Fontaine, J.R.J., Soriano, C.: The GRID meets the Wheel: Assessing emotional feeling via self-report. Components of emotional meaning: A sourcebook pp. 281–298 (2013)
27. Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Lucas, R.E., Agrawal, M., Park, G.J., Lakshmikanth, S.K., Jha, S., Seligman, M.E., Ungar, L.: Characterizing geographic variation in well-being using tweets. In: Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM). pp. 583–591 (2013)
28. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Computational linguistics 37(2), 267–307 (2011)
29. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A.: Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology 61(12), 2544–2558 (2010)
30. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology–Volume 1 (NAACL-HLT). pp. 173–180. ACL (2003)
31. Wang, W., Chen, L., Thirunarayan, K., Sheth, A.P.: Harnessing Twitter "Big data" for automatic emotion identification. In: Proceedings of the International Confernece on Social Computing (SocialCom). pp. 587–592. IEEE (2012)
32. Yang, C., Lin, K.H.Y., Chen, H.H.: Building emotion lexicon from weblog corpora. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. pp. 133–136. ACL (2007)
33. Zhu, X., Guo, H., Mohammad, S., Kiritchenko, S.: An empirical study on the effect of negation words on sentiment. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL). pp. 304–313. ACL (2014)