CONSTRUCTION AND ANALYSIS OF AN EMOTION PROPOSITION STORE

SEBASTIAN RUDER

Bachelor's thesis

Institute for Computational Linguistics
University of Heidelberg

July 2015

Sebastian Ruder: *Construction and Analysis of an Emotion Proposition Store,* Bachelor's thesis, © July 2015

supervisor:
Prof. Anette Frank

location:
Heidelberg

ABSTRACT

Emotion detection, a subtask of sentiment analysis, focuses on identifying emotions in text. Knowledge about emotion expressions and emotions associated with events can be beneficial in applications such as facilitating interaction with AI agents, understanding customer's feeling towards products, assisting governments in recognizing growing anger or fear and pre-empting mass hysterias. For this purpose, we design and evaluate patterns that are frequent and clearly associated with an emotion. These patterns can be used as-is to extract tuples of emotion holders and causes from the web as well as from special domain corpora. Using these patterns, we acquire more than 1,700,000 propositions from the Annotated Gigaword news corpus, filter, and generalize them by employing co-reference resolution and named-entity recognition. These propositions contain information about the emotion, the emotion holder, and the cause of said emotion. We store these propositions in an Emotion Proposition Store, which we make available to the research community. We analyse and evaluate them to gain further understanding about emotions in news text as well as the capabilities of the resource. Distributional analysis and topic modelling allow us to determine ambiguous concepts, underlying themes, as well as single-word and compound expressions that are highly associated with an emotion, which we make available in emotion lists that can be used as an emotion lexicon for the news domain.

ACKNOWLEDGEMENTS

I wish to thank the staff and faculty of the Institute of Computational Linguistics at the University of Heidelberg for being a stimulating learning environment, for putting me on the path of natural language processing, and for endowing me with the knowledge to further pursue this path. I wish to thank my fellow undergraduate students who have accompanied me on this path; I particularly want to express my gratitude to Julius Steen and Devon Fritz for their help in the annotation task.

Above all, I would like to thank my supervisor Prof. Dr. Anette Frank for her guidance, her limitless patience in light of my ceaseless requests for references, her flexibility in communicating during my semester abroad, and her good advise. I wish to thank Dr. Alexis Palmer, whose comments introduced coherence to my ramblings. I also wish to thank Dr. Michael Herweg for his conversations and seminars that have shaped my fledgling interests in natural language processing.

I wish to thank Dr. Carl Vogel of Trinity College, Dublin for his guidance and advise during my semester abroad. I also want to thank my colleagues in the Linguistic Engineering Team at Microsoft, James Cogley and Declan Groves, for their guidance during my internship and their help with annotation.

Finally, I wish to thank my family, my friends, and particularly my girlfriend Jessica for their endurance and patience. I wish to dedicate this thesis to the memory of my mother and wish to thank her for making me who I am.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

Categorizing sentences according to their sentiment (or opinion) is an important first step for many downstream applications for natural language processing (NLP) such as dialogue systems [40], text-to-speech systems [4], media response analyses [42], and a host of others. Sentiment analysis is usually restricted to a polar classification: positive – negative with a non-sentiment-bearing neutral class. This binary classification falls short in light of the variety and multifacetedness of human emotions. Emotions are innate to humans, so much so that the inability to show emotions is considered a clear indicator of a societal outlier. They transcend cultures and facial expressions for basic emotions are identical even between cultures with no anterior contact [12]. Recent work in sentiment analysis also includes the subtask of emotion analysis or emotion detection, which focuses on identifying emotions in text typically using a set of emotion categories suggested by psychological research (e.g. the eight classes – joy, trust, fear, surprise, sadness, disgust, anger, anticipation – of Plutchik's wheel of emotions [38]) (cf. [29]). Identifying emotions in text can be beneficial in e.g., facilitating interaction with AI agents, helping companies understand people's feelings towards their products, assisting governments in recognizing growing anger or fear towards an event, pre-empting mass hysterias, or helping media companies understand people's emotional reaction towards controversial issues.

The main contributions of this thesis to the branch of emotion detection are to:

1. Design and evaluate patterns that are frequent and clearly associated with an emotion. These patterns can be used as-is to extract tuples of emotion holders and causes from the web as well as from special domain corpora.
2. Acquire more than 1,700,000 propositions from the Annotated Gigaword news corpus [32] using these patterns, filter, and generalize them by employing co-reference resolution and named-entity recognition (NER). These propositions contain information about the emotion, the emotion holder, and the cause of said emotion.
3. Store these propositions in an emotion proposition store, which we make available to the research community. Two sample entries are depicted in figure 1.
4. Analyse and evaluate them to gain further understanding about emotions in news text as well as the capabilities of the resource.

```
NYT_ENG_19960601.0010/1 trust rely on Japan/LOCATION economic
    clout     [its/PRP$, economic/JJ, clout/NN]
NYT_ENG_19960601.0010/9 anticipation expect Japanese    win
    right  [to/TO, win/VB, the/DT, right/NN, to/TO, host/VB,
    the/DT, tournament/NN]
```

Figure 1: Proposition examples

> Distributional analysis allows us to determine ambiguous concepts, underlying themes, as well as single-word and compound expressions that are highly associated with an emotion, which we make available in emotion lists that can be used as an emotion lexicon for the news domain.

Past research has shown that both the automatic categorization of emotion (cf. [14]) as well as the identification of the holder and the cause of an emotion is difficult (cf. [9]). Additionally, linguistic triggers of emotions are diverse and can belong to different word classes: nouns like *disaster*, *prize*, or *crash*; adjectives like *mind-numbing*, *yucky*, or *moist*; and verbs like *cheat*, *praise*, or *betray* are all equally able to evoke certain emotions. These issues render a pattern-based approach attractive that uses clearly categorized linguistic cues that also clearly indicate the emotion holder and cause to automatically label emotion-triggering concepts and events. We select frequent and clearly emotion-indicating predicators as patterns such as *X be angry about NP / that S*, *X fears that S*, *X was surprised that S*, *X (absolutely) hates NP / S*, etc.

The pattern-based approach we take can also be applied to automatically extract a general-domain compositional emotion lexicon from the web (using a web mining approach as in [22]) as well as be extended to other special domain corpora, to learn about what concepts and entities people love, hate, fear, or wish for, and where their emotions differ. In comparisons to other approaches like manual curation or crowd-funding [29], it requires only a minimum amount of up-front time investment, as our compiled patterns can be used as-is.

We focus on news for this thesis, as many sentiment applications concentrate on this domain, with real-world use cases such as alerting traders and predicting the stock market [7] and as researchers have shown that domain-specific knowledge generally outperforms general-domain knowledge [3]. In consequence, we use the Annotated Gigaword [32] corpus, which adds off-the-shelf annotations to the biggest static English news corpus in existence.

Using this pattern-based approach, we are able to derive from a sentence such as *He was angry because prices had gone up* that *prices had gone up* is an event that triggers an emotion of anger in the subject of *be angry*. However, given the cause of an emotion, it is not evident which part is most relevant for its association with the emotion. Emotion (just as sentiment) is a compositional phenomenon: Neither *prize* nor *go up* are by themselves emotion-triggering words, but compositional expressions such as *the prize went up* or *rising prizes* are prone to trigger emotions. We derive this from the distributional properties of the harvested propositions in case the individual parts often appear jointly with the emotion, i.e. yield a high association score.

While emotion lexica, i.e. lexical resources containing sentiment-labeled expressions, have been of frequent use in sentiment analysis, resources for emotion detection are still rare. Our emotion proposition store will help to alleviate this deficit: The lists of emotion-associated words obtained through distributional analysis can be used in the same vein as sentiment or emotion lexicons. Moreover, as these lists consist of both unigrams and bigrams, which are categorized in concepts (nominal phrases) and events (clauses), they provide more context and thus more potential than other emotion lexicons containing unigrams. Finally, the store's tuples of contextualized emotion holders and causes can be used in other tasks, e.g. for inference. For instance, making sense of a sentence like *(I think) people are happy because Chavez has fallen* [10] requires knowledge that $E : \texttt{fall}(X)$ is an event that is bad for X, while $\texttt{happy\_about}(Y, E)$ states that Y is happy about event E. Through inference, we are able to derive that Y has a negative attitude toward X. Drawing such inferences presupposes knowledge contained in our emotion proposition store about positive and negative emotion expressions ranging over propositions and positively and negatively perceived events.

The structure of this thesis is as follows: In chapter 2 we will examine past research relevant for our approach, including methods of emotion detection, psychologically-grounded typologies of emotion classification, and sentiment lexica. This serves the dual purpose of introducing key concepts of emotion detection and sentiment analysis as well as touching on some of the tools and resources to be applied in our research.

In chapter 3, we present the chosen corpus in more detail. Moreover, we will introduce premises that have shaped the compilation of the emotion-bearing patterns, provide exemplary patterns, and give an overview of their sources. We will discuss the evaluation of the patterns and the final list of patterns that we used for extraction, at which point we will detail the extraction steps and introduce the chosen representation format.

Subsequently, in chapter 4, we will analyse and evaluate the results of the extractions. We will compare two association metrics, the salience of single and compound emotion-associated expressions, ambiguous concepts, as well as nominal and causal causes. We evaluate the highest-ranking emotive expressions against an emotion lexicon as well as against human-annotated data. Moreover, we use an extension of Latent Dirichlet Allocation, a method to discover latent semantic topics in text, in order to explore the underlying themes that are associated with emotions.

Finally, in chapter 5, we will both look ahead, outlining promising paths for future research, and look back, summarizing our findings and offering a conclusion.

2 FOUNDATIONS AND RELATED WORK

Research on sentiment analysis has been vast and there are excellent summaries (cf. [34], [21], [25]). In the following, we will focus on work related to emotion detection and sentiment analysis as well as on idiosyncrasies of emotions. In section 2.1, we will discuss the compositionality of emotions. Subsequently, in section 2.2, we will outline research on emotive events, whereas in section 2.3, we will describe pattern-based mining approaches for emotion detection. In section 2.4, we will give an overview of semantic roles relevant in the context of emotions, while in section 2.5, we will deal with psychologically-grounded typologies of emotion classification. Finally, in section 2.6, we will introduce popular sentiment and emotion lexicons. To avoid ambiguity, we will henceforth define sentiment as being positive or negative, while an emotion possesses more complex dimensions.

2.1 COMPOSITIONALITY

Sentiment and emotion can both be considered as compositional phenomenons. Their compositionality, however, is of a different nature: Sentiment clearly adheres to Montague's underlying *principle of compositionality* and the notion of sentiment compositionality has been successfully applied in state-of-the-art approaches such as recursive neural tensor networks [44], which conceive the overall sentiment of a compound expression as the sum of the constituent polarities. In turn, compositionality for emotions is less clear: Two unemotive terms can – combined – evoke an emotion, e.g. *the blind* and *sees* are both inherently neutral terms; concatenated, the compound *the blind sees* – seeming either miraculous or oxymoronic – evokes an emotion of supreme joy. This emotion, however, cannot be calculated merely as the sum of its parts and can only be accurately predicted by making sense of the underlying world knowledge.

To illustrate the difficulty of predicting compositionality for emotion versus sentiment, we cite the work of Borth et al.: They construct a visual sentiment ontology containing more than 3,000 adjective-noun pairs (ANP). They obtain tags from Flickr and Youtube videos that are associated with any of Plutchik's emotions, rank them based on frequency, and determine their sentiment. This way, they retrieve 1,187 nouns (576 positive/negative, 611 neutral) and 268 positive and negative adjectives. They pair the adjectives and nouns to obtain ANPs such as *beautiful flower* or *disgusting food*. They are able to calculate the sentiment of the compound expression simply by summing the sentiment values of the adjective and the noun, assigning a higher weight to the adjective in case of opposite sentiment values to account for ANPs like *abused child*. As emotion does not lend itself to easy compositionality, they have to fall back on additional information taken from Flickr meta-data to determine the emotion of the ANP.

2.2 EMOTIVE EVENTS

In the above case, emotion is predicted for static entities. Another focus of research in emotion detection are events. Evidently, predicting emotion for events is harder than for nominal phrases, as modifiers contain a lot of meaningful information. As the above authors realized, adjectives like *abused*, *disgusting*, *happy*, etc. almost always propagate their respective emotion to the whole nominal phrase. As we have alluded to in chapter 1, knowledge about the emotive connotation of an event is valuable for inferences.

Deng et al. try to capture such implicit attitudes towards events by annotating benefactive and malefactive events, i.e. events that negatively or positively affect entities. They let two annotators annotate a total of 725 sentences and report 0.69 as overlapping agreement between event spans and good agreement (0.87 and 0.97) for agents and objects. Their annotation scheme, however, raises doubts, as neutral events such as *She bakes a cake* are likewise labelled as benefactive or malefactive.[1]

Lu et al. leverage world knowledge to detect emotions on an event level. Whereas we use association metrics derived from emotive contexts, their main insight is that the *common mutual actions* between two event participants are a major cue for detecting emotions. For instance, a snake in general performs undesirable actions towards a girl, which would give the event *a girl sees a snake* a negative emotion.

Given an event, they also account for active and passive. However, they don't match the patterns against an existing corpus as we do, but mine data from the web to generate mutual action histograms. They compare these to the histograms of a set of manually emotion-labelled events between reference pairs and return the emotion of the most similar histogram. They evaluate their system on 600 events manually labelled by ten annotators with a positive, negative, or neutral sentiment, omitting finer granularity. Their proposed emotion detection system achieves an accuracy of 81.0% on events where a majority of six annotators agreed.

1 *Cake* would be the object of the benefactive event in this example.

2.3 PATTERN-BASED MINING FOR EMOTION DETECTION

Similarly, Pantel and Pennacchiotti leverage the web to automatically extract instances using patterns. Their method starts off with a set of seed instances, from which they induce and then select new patterns. In contrast to most unsupervised approaches, they not only use *reliable* patterns (high precision, low recall), but also leverage generic patterns. They mitigate their imprecision by measuring pattern reliability based on instantiations of reliable patterns, which enables them to separate correct from incorrect instances. They evaluate their system on seven relations and show that without generic patterns, their system outperforms comparable systems in precision, while generic patterns improve its recall significantly, e.g. from a relative recall of 1.00 to 6.66 for the `is-a` relation in the CHEM corpus, while precision decreases less sharply from 85.0% to 76.0%.

While their approach is helpful for gathering knowledge to ontologize `is-a` and `part-of` relationships, the use of generic patterns for emotion detecion, which is already highly ambiguous, would introduce more noise and consequently rather hurt than improve performance. Furthermore, the difficulty of finding seed instances that will generate reliable patterns is amplified for emotions: *I :: spiders*, *I :: darkness*, *I :: violence*, and *I :: Bin Laden* might be promising seed instance to induce the pattern *X hate Y*; they could also as easily be associated with *X am scared of Y*, *X object to Y*, etc. That is, while it is relatively straightforward to distil the essence of an `is-a` relationship via a set of seed instances, doing this for emotions is less easy and will invariably introduce a selection bias. Insofar, starting with a set of seed patterns and expanding them would be the preferred approach for emotion detection. We will leave this for future work.

Fellbaum and Mathieu in turn use search query patterns to rank verb classes. They manually construct classes corresponding to the emotions surprise, fear, and annoyance and rank them by intensity of emotion by mining the lexical-semantic patterns *X (perhaps) even Y*; *X, not to say Y*; and *If not X then Y* from the web. Their results indicate that some verbs such as *intimidate*, *scare*, and *alarm*, while being similar, are not members of the same broader *frighten* class. Example 1 shows their ordering of *surprise* verbs.

astonish > surprise > amaze > astound > strike > stun > floor
> dumbfound > flabbergast > stupefy

(1)

While seed instances and patterns can be emotionally ambiguous, the micro-blogging platform Twitter provides a more salient indicator of emotions: hashtags. Mohammad as well as Qadir and Riloff use hashtags to extract emotion-evoking tweets. The former use emotions as hashtags, e.g. #*anger*, #*joy*, etc. to construct a Twitter emotion corpus by extracting 21,000 tweets containing these hashtags at the end of the message from the Twitter Search API[2], with the joy emotion being prevalent in their corpus. They use an ensemble of binary support vector machine (SVM) classifiers to predict the emotion hashtag of a tweet, achieving an F-score of 49.9 across all emotions, with disgust achieving the lowest (18.7) and joy the highest (62.4) scores.

Qadir and Riloff in turn use a bootstrapping approach to learn new hashtags and hashtag patterns and harvest emotion phrases from these. They start by collecting 323,000 tweets containing seed hashtags like #*loveyou*, #*sweetheart*, and #*bff* for five emotions. They train a logistic regression classifier for each emotion, using tweets containing the seed hashtags as positive instances, with negative instances consisting of randomly selected tweets from a pool of unlabelled tweets. Applying the classifier to the pool of unlabelled tweets produces new emotion-labelled tweets, from which they retrieve the ten highest scoring emotion-indicating hashtags by computing the average probability assigned by the classifier. They also learn hashtag pattern (e.g. #*missedyoutoomuch* → #*missedyou**) and emotion phrases (e.g. #*lovemylife* → *love my life*) by representing hashtags in a trie data structure, applying the classifier to tweets matching the patterns and phrases, and selecting the ten patterns and phrases with the highest average probability respectively. They achieve the best performance (51-66 F-score across emotions) on a manually annotated corpus of 5,500 tweets with a hybrid approach of an SVM with n-gram features enhanced with the phrase-based classifier and the classifiers of their learned hashtags and hashtag patterns.

2.4 SEMANTIC ROLES IN THE CONTEXT OF EMOTIONS

In the context of emotions, not only the nature of the emotional state, but also the person experiencing the emotion and the cause of the emotion are relevant. Semantic role labeling is the task in NLP that aims to identify these semantic arguments of the predicate and classify them into their specific roles. FrameNet [2], a linguistic resource that stores semantic role information in so-called frames related to certain contexts, lists the roles depicted in table 1 as core roles for its EMOTIONS frame.

As the event is most often given, most research focuses on identifying the Experiencer and the Stimulus of the cause. In the following, we will refer to the Experiencer and the emotion holder as well as the Stimulus and the cause interchangeably.

Das and Bandyopadhyay construct two models to identify the holder of an emotion. The baseline model leverages the *subject* dependency of emotional sentences, while an unsupervised syntax-based model

2 https://dev.twitter.com/rest/public/search

Role	Description
Event	The occasion or happening that Experiencers in a certain emotional state participate in.
Experiencer	The person or sentient entity that experiences or feels the emotions.
Expressor	The body part, gesture, or other expression of the Experiencer that reflects his or her emotional state.
State	The abstract noun that describes a more lasting experience by the Experiencer.
Stimulus	The person, event, or state of affairs that evokes the emotional response in the Experiencer.
Topic	The general area in which the emotion occurs. It indicates a range of possible Stimulus.

Table 1: Core semantic roles in FrameNet's EMOTION frame

depends on the argument structure of emotional verbs, which they acquire in two ways: (a) from dependencies and (b) from the part-of-speech tagged and chunked sentences. They compare the sentence argument structure against frame syntaxes of 942 emotional verbs extracted from VerbNet [19], a linguistic resource that concentrates on verbs. They report an F-Score of 64.83% for the baseline and of 66.98% and 62.39% for the syntax-based model with argument structures acquired using dependencies and parts-of-speech respectively.

Mohammad et al. frame the task of identifying the Experiencer, State, and Stimulus of an emotion as a classification task. They train their classifier on a dataset of 2012 US presidential election tweets with annotations crowd-sourced to Amazon's Mechanical Turk[3] and report an accuracy of 56.84 for detecting emotions and an F-score of 58.30 for detecting the stimulus. As their tweets deal with a narrowly defined domain, they are able to select the stimulus from a set of pre-chosen entities, which is successful in this narrow domain, but fails given an open set of potential stimulus.

Balahur et al. in contrast don't use information extraction (IE) techniques but base their approach on their knowledge base EmotiNet. As their knowledge base is focused on the family domain, they only consider that domain, where their approach based on lexical resources extended with EmotiNet outperforms an SVM algorithm trained on more data achieving an average F-Score of 45.27% across emotions. They remark that world knowledge is particularly valuable in cases where no emotion-bearing word, e.g. *happy* is present.

3 https://www.mturk.com/mturk/welcome

2.5 TYPOLOGIES OF EMOTION CLASSIFICATION

Psychologists have proposed a number of different theories for classifying emotions, e.g. distinguishing between instinctual and cognitive emotions. As emotions are inherently subjective, none has been universally accepted. Two classification theories have been frequently used by the research community: Ekman's and Plutchik's. Ekman proposes a theory with six basic emotions: anger, disgust, fear, joy, sadness, and surprise. According to Plutchik, there are two additional ones: anticipation and trust. He depicts them in a wheel, shown in figure 2, where the intensity of the emotion increases with the proximity to the center, and organizes them in four opposing pairs: joy–sadness, trust–disgust, fear–anger, surprise–anticipation.

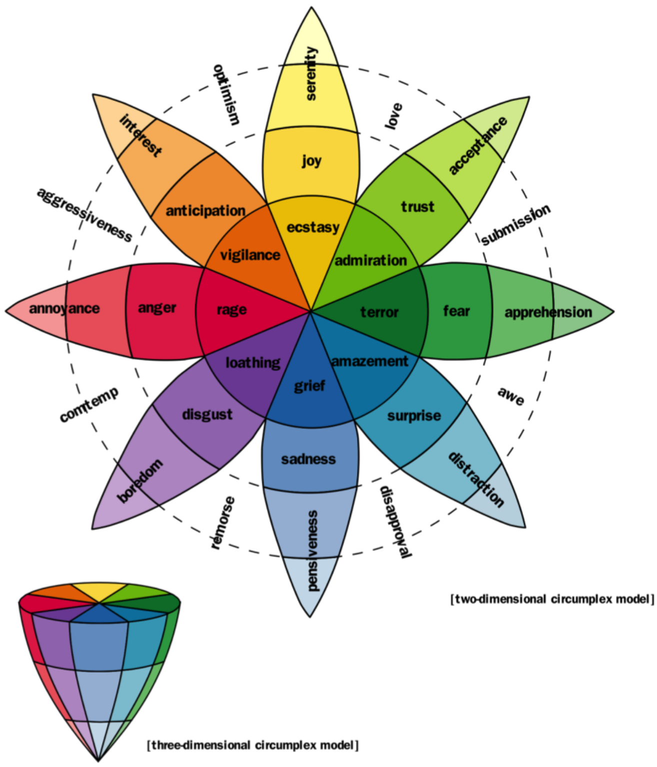

Figure 2: Plutchik's emotion wheel

We adopted Plutchik's emotion classification for the following reasons: (a) it is clearly founded in psychological research; (b) it is a superset of Ekman's six basic emotions; and (c) its use by other researchers (cf. [29], [8], [27]) increases the comparability and transferability of our work.

2.6 SENTIMENT AND EMOTION LEXICONS

We will receive concepts, entities, and events that are associated with certain emotions via analysis of our extracted propositions. We will store these concepts in emotion-labelled lists that can be used in the same vein as existing lexicons. Sentiment lexicons in particular, as reliable sources of knowledge about the sentiment of concepts, have been pervasive in sentiment analysis. In the following, we will give an overview of the most commonly used ones.

The General Inquirer [46], the first sentiment lexicon to our knowledge, was created in 1961. It contains 11,788 words labeled with 182 categories of word tags, some of which are relevant to emotions: Among them are 1045 positive words, with a subset tagged for indicating affiliation or supportiveness, and 1160 negative words, a subset of which is tagged for indicating an attitude or concern with hostility or aggressiveness. 168 words are tagged with PLEASURE indicating the enjoyment of a feeling; 254 PAIN words indicate suffering, lack of confidence, or commitment; 49 FEEL words describe particular feelings, including gratitude, apathy, and optimism; 166 AROUSAL words indicate excitation (aside from pleasures or pains); finally 311 EMOT words relate to different emotions such as *admiration*, *annoyance*, etc. The General Inquirer was key in early experiments such as differentiating fake suicide notes from real ones. Since then, it has been used for a plethora of other applications, particularly in sentiment analysis, e.g. for contextualizing polarity [49].

Two more recent sentiment lexicons are WordNet-Affect [47] and SentiWordNet [13]. Both are built on top of the popular natural language processing resource WordNet [26]. WordNet is a lexical database that organizes nouns, verbs, and adjectives into *synsets* that pertain to underlying lexical concepts. WordNet-Affect annotates WordNet synsets representing affective concepts, whereas SentiWordNet assigns objectivity, positivity, and negativity scores to WordNet synsets.

The manual compilation of a sentiment lexicon can be a long and tedious process. Recently, researchers have looked for other methods: Mohammad and Turney use crowd-sourcing to build a sentiment lexicon containing about 10,000 entries. They take their terms primarily from the *Macquarie Thesaurus*, as well as the General Inquirer and WordNet-Affect and crowd-source the annotations using Amazon Mechanical Turk. They use Plutchik's eight emotion classes as well as positive and negative sentiment as binary labels. Mohammad and Yang follow the same approach to build the NRC Emotion Lexicon, containing about 14,000 word types, using *Roget's Thesaurus* instead of the *Macquarie Thesaurus*. They use it to quantitatively compare the emotion words in love and hate mail, as well as between genders and use it for other applications, e.g. as features for training the SVM classifier on election tweets mentioned above [31].

Staiano and Guerini finally use a different form of crowd-sourced information to create a sentiment lexicon: `rappler.com` lets users vote on the mood of a news story. The authors initially build a document-by-emotion matrix, listing each story with the voting percentages for all eight moods, and a term-by-document matrix using different frequency measures. Matrix-multiplication and normalization produces a final word-by-emotion matrix containing 37,000 entries.

3 PATTERN COMPILATION AND EXTRACTION

In this chapter, we describe our approach to building patterns and using them to extract emotion-bearing propositions from a large corpus. Firstly, we introduce our corpus in section 3.1. In section 3.2, we detail the principles that guided the design of a methodology for pattern-based extraction. In section 3.3, we give examples of patterns and subsequently, in section 3.4, we describe the various sources on which we base the design of our patterns. Following, in section 3.5, we detail the compilation of regular expression patterns from our collected pattern templates and afterwards, in section 3.6, we describe the annotation task we use to reduce our collected patterns to a final list of patterns that are clearly indicative of one emotion and that we present in 3.7. In section 3.8, we describe the actual pattern-based extraction process and finally, in section 3.9, we give an overview of the representation format we have chosen for the extracted emotion-bearing propositions.

3.1 CORPUS

Genre and domain clearly play an important role in the context of emotions and must be taken into consideration for the selection of an adequate corpus. Some genres are designed to elicit emotions, while others are geared to other functions. Insofar, the domain is probably the factor with the most impact on the results. Mohammad show that fairy tales and certain book genres are rich in emotion.

In the news domain, sentiment analysis is already being widely applied; the application of emotion detection to this domain in consequence constitutes a logical, next step. Additionally, emotion detection in news poses an interesting challenge, as emotions in news are expected to be less overt than in emotional literature. Furthermore, obtained insights and results will be not only relevant in the context of a particular oeuvre, but on a national or international scale. Finally, the English Gigaword corpus is a resource that provides a wealth of data that is not as readily available for other domains.

# tokens	**# documents**
4,032,686,000	9,876,086

Table 2: Number of tokens and documents for English Gigaword v.5

Not only is its latest addition, Gigaword v.5 [37], containing almost 10 million documents from seven news outlets, with a total of more than 4-billion words (cf. table 2), the largest static corpus of English news documents in existence; but also Napoles et al. recently made an annotated version of it available that facilitates processing. They provide the following pre-processing all of which we are using:

A. tokenized and segmented sentences,

B. Treebank-style constituent parse trees,

C. syntactic dependency trees,

D. named entities, and

E. in-document coreference chains.

The Annotated Gigaword has already been used for approximate search, e.g. to find relations similar to *X dived Y*. As our patterns are of a similar format, it is ideal for our approach.

3.2 MOTIVATIONS

In this section, we present the principles that motivated the design of our emotion-bearing patterns in order to extract emotion holders and causes from the corpus. They are the following:

A. Avoidance of ambiguity: patterns should be unambiguous and clearly indicate one emotion.

B. Extractability of emotion holder and cause: patterns should clearly indicate the holder and the cause of an emotion.

C. Diversity of causes: patterns should account for a diverse set of causes such as entities, concepts, events, actions, and conditions.

3.2.1 *Avoidance of ambiguity*

A clear concern for the compilation of patterns is ambiguity: As emotions are ambiguous, many expressions can be indicative of more than one emotion, e.g. *anxious* can both refer to a state of eagerness or worry. Of 165 anger terms in WordNet-Affect, annotators have labelled 53% as well with disgust [29], showing a clear overlap between these two emotions.

We seek to pre-empt this ambiguity in two ways: (a) We only select expressions that are listed as pertaining to one emotion in one of our sources as detailed in section 3.4; and (b) measure agreement between three annotators on the selected expressions, choosing only those on which a majority agrees. We could have included multiple emotions for each pattern, but decided against it, as we ultimately wanted to obtain expressions that are clearly associated with one emotion.

Furthermore, Fellbaum and Mathieu note that emotion verbs can overlap with cognition verbs: *shock* can evoke both a judgement or an emotion. We intend to overcome this issue by (a) selecting patterns whose dominant sense is emotive; and (b) by investigating propositions extracted using our patterns. An analysis of a representative sample will surface erroneous patterns that lead to unemotive contexts in the news domain, which we will be able to retroactively exclude.

Our overarching goal is the harvesting of emotion-evoking expressions. Insofar, we avoid using complex sentence structures and omit error-prone concepts like proverbs, multiple embeddings, or idiosyncracies. Instead, our patterns focus on clear-cut cases and structures that can be easily disambiguated, including patterns like *X fear Y*, *X be angry about Y*, *X be happy about Y*, etc. that do not rely on context for disambiguation and are clearly emotion-bearing. Initial experiments have shown that part-of-speech labelling is important, as e.g. *trust* and *fear* can function both as verbs and nouns. In addition, lemma representation is valuable on a representational level to guarantee generalizable results.

3.2.2 *Extractability of emotion holder and cause*

As we have outlined, we seek to show which expressions or entities are associated with or evoke certain emotions. Mohammad and Turney show that there is a slight distinction between *associated* and *evokes*, whereupon *associated* produces moderately higher agreement numbers between annotators when used in the annotation task description. As this agreement is not relevant for the annotation of our patterns, we will treat *associated* and *evokes* equivalently in the following.

In order to show association, we need to be able to extract both the holder and the cause of an emotion. Das and Bandyopadhyay's syntax-based model, which we described earlier, that acquires argument structures via dependencies and matches them against VerbNet frames has an F-score of 66.98%. In order to create a reliable resource, our patterns need to be more accurate.

Insofar, one necessary condition for the selection of a pattern is the presence of both the emotion holder and the cause. For instance, we would not like to choose simply *be happy* as a pattern since *I am happy* does not indicate the cause of happiness. *Be happy that* would be a more suitable pattern, as *I am happy that you came* provides more information as to the stimulus of the state of being happy. This leads us to mainly employ transitive constructions. Fellbaum and Mathieu only include Experiencer psych verbs in their emotion classes. In the transitive construction of these verbs, the subject refers to the cause, while the object expresses the Experiencer, as depicted in example 2a.

Thus they include verbs like *scare* and *intimidate*, but exclude verbs like *fear*, whose subject in a transitive construction expresses the Experiencer, as in example 2b.

The dog frightens the man.
CAUSE EXPERIENCER (2a)

The man fears the dog.
EXPERIENCER CAUSE (2b)

Verbs like *admire*, *deplore*, *abhor*, etc. possess the same property[1] and are clearly emotion-bearing as well as indicative of emotion holder and cause. In order to adequately capture this property, we include information along with each pattern indicating if the order of subject–EXPERIENCER, object–CAUSE depicted in example 2b is reversed.

3.2.3 *Diversity of causes*

If we only capture objects as causes, we will clearly miss out on a lot of variety that is inherent to emotions. Emotions can equally be evoked by a concept, entity, event, situation, or condition, with events being a particular focus of research as we described in section 2.2. Besides capturing nominal phrases, we thus also want to collect patterns that indicate emotions whose stimulus is an act, event, or condition encompassed in a sub-clause as can be seen in example 3.

I fear that my company will go bankrupt.
EXPERIENCER CAUSE (3)

The types of causes we thus expect are nominal phrases or clauses. These are labeled as NP and S/SBAR respectively in phrase-structure grammars. In the Penn Treebank tags – the tag set that is used by the Stanford parser in the Annotated Gigaword corpus – S is a simple declarative clause, SBAR is a declarative clause introduced by a subordinating conjunction, and NP is a noun phrase. Thus, in the sentence in example 4, *after she ate the cake* is an SBAR introduced by the subordinate conjunction *after*, while *she ate the cake* and the whole sentence are declarative clauses.

After she ate the cake, she visited him. (4)

If a predicate can occur with both a nominal phrase and a sub-clause, we will create a separate pattern for each cause type and add information to indicate the type.

1 They are all contained in VerbNet's ADMIRE class, as VerbNet groups verbs with the same syntactic alternation behaviour together.

3.3 PATTERNS

So far, we have stated the motivation for the design of our patterns without making abundantly clear what form they will take. For disambiguation purposes, we treat a pattern as an emotion-bearing predicate, e.g. *be happy that*, while a pattern template contains the pattern along with its emotion and two binary flags specifying meta-data:

A. The first one indicates if the cause of the emotion is a nominal phrase or a clause or sub-clause. It takes the values of NP (nominal phrase) and S (clause or sub-clause).
B. The second flag specifies if the order of subject–EXPERIENCER, object–CAUSE as described in section 3.2.2 is reversed. It takes as values TRUE – if the order is reversed – and FALSE – if the order is maintained.

If a verb or an adjective has been listed in one of our sources specified in section 3.4 with an emotion, we adopt that emotion, given it is among Plutchik's eight emotion classes; if not, we manually label it. We list the pattern in lemma form along with its Penn Treebank part-of-speech tags. We list verbs with the generic tag VERB to represent the eight verb part-of-speech tags in the Penn Treebank.

Two pattern templates are given in example 5:

fear scare/Verb NP true
joy be/Verb RB happy/JJ that/IN S false (5)

In the first pattern template, the pattern *scare* has the emotion fear, it takes as direct object a nominal phrase, and its subject has the semantic role of the cause (as indicated by the reversed order). In the second pattern template, joy is the emotion of the pattern *be happy that*, which takes a clause as its complement, while the subject is the emotion holder. RB indicates that we allow *happy* to be modified by any adverb.

3.4 PATTERN SOURCES

We make use of the selection of emotion lexicons available as well as resources mentioned in previous works to collect emotion-bearing patterns. We use three categories of sources: (a) dictionaries and thesauri; (b) sentiment lexicons as well as past research; and (c) general-purpose natural language processing resources.

3.4.1 *Dictionaries and thesauri*

We use the emotion's definition in the Oxford English Dictionary to generate an initial list of patterns. The Oxford English Dictionary is widely acknowledged to be the most comprehensive and authoritative record of the English language, which makes it an apt starting point for our endeavour. For instance, it defines joy as:

A vivid emotion of pleasure arising from a sense of well-being or satisfaction; the feeling or state of being highly pleased or delighted; exultation of spirit; gladness, delight.

From this definition, we are able to extract the patterns *satisfy*, *please*, and *delight*. In total, we retrieve 24 initial patterns from the Oxford English Dictionary.

We then use Merriam-Webster's Dictionary and Roget's Thesaurus – which was used to produce target terms by Mohammad and Yang – to retrieve synonyms for those verbs. As Roget's Thesaurus has a relevance ranking, we choose only those synonyms with the highest relevance. For joy, we retrieve in Merriam-Webster's Dictionary patterns such as *rejoice*, *jubilate*, or *triumph*, while we obtain from Roget's Thesaurus patterns such as *exult*, *revel*, *make happy*, *amuse*, *charm*, *enchant*, and many others. Merriam-Webster's Dictionary in total produces 51 patterns, while Roget's Thesaurus produces 101 patterns, confirming the quality of this resource and justifying its frequent application in lexically-based approaches.

3.4.2 *Sentiment lexicons and past research*

3.4.2.1 *Harvard General Inquirer*

While the Harvard General Inquirer [46] has found use in a plethora of applications and contains a multitude of useful annotations, its EMOT category does not distinguish emotions on the finely granular level that we require for emotion detection. Furthermore, most of its words are nouns that cannot be used in our predicates. For these reasons, we are not using it in this project.

3.4.2.2 *NRC Word-Emotion Association Lexicon*

As has already been mentioned, the NRC Word-Emotion Association Lexicon (also called EmoLex) was developed by Mohammad and Yang using crowd-sourcing. While it also employs Plutchik's eight emotion classes, annotations can be very subtle and rather mirror underlying perception tendencies than what would be helpful in surfacing the holder and the cause of an emotion, as can be seen in example 6, where *abacus* is associated with the emotion trust. In consequence, we are not able to retrieve any clearly emotion-indicating patterns from it.

abacus trust 1 (6)

However, in the same vein, its entries, which evoke or are associated with emotions, are similar to the kind of entries that we would want to have in our proposition store. Insofar, we will use it for a partial evaluation of our proposition store in section 4.3.1.

3.4.2.3 *Emotion verb classes*

The manually constructed emotion verb classes by Fellbaum and Mathieu fulfil the requirements for our patterns: They clearly indicate an emotion and the emotion holder and cause can be extracted using the semantic roles.

We incorporate the following ten, five, and nine members as emotion verbs for Plutchik's emotions surprise, fear, and anger, respectively, as anger is a degree of annoyance; for surprise: *astonish*, *surprise*, *amaze*, *astound*, *strike*, *stun*, *floor*, *dumbfound*, *flabbergast*, and *stupefy*; for fear: *intimidate*, *scare*, *frighten*, *alarm*, and *terrify*; for anger: *irk*, *nettle*, *irritate*, *annoy*, *anger*, *exasperate*, *infuriate*, *enrage*, and *incense*.

3.4.2.4 *Adjectives*

As we have seen, linguistic triggers for emotions are diverse; not only verbs, but also adjectives are apt in capturing emotions. In WordNet, nouns and verbs are clustered in synsets and supersenses. Tsvetkov et al. induce supersenses for adjectives – taking GermaNet's guidelines[2] as inspiration – in order to create a taxonomy for adjectives.

They build a weakly supervised classifier that labels adjective types (irrespective of context), which they train on a small set of seed examples, some of them translations of GermaNet. We take the 126 adjectives that pertain to FEELING and manually label them with Plutchik's eight emotion classes.

They also released 7511 WordNet adjectives tagged by their classifier with an adjective type vector. Of these, it labels 920 adjectives with the adjective type FEELING. As the classifier only has an accuracy of 54%, they consist of many erroneous instances like *unknown*, *unreal*, etc. rendering this set of adjectives unusable for our purposes.

3.4.3 *General-purpose NLP resources*

The troika of resources, WordNet, VerbNet, and FrameNet, has seen frequent use in NLP applications leading us to investigate the use of the three for our purposes.

3.4.3.1 *VerbNet*

VerbNet [19] focuses on the syntax and semantics of verbs, clustering them in semantic groups pertaining to their occurrence in syntactic al-

2 http://www.sfs.uni-tuebingen.de/lsd/adjectives.shtml

ternations. While it is helpful in identifying the holder and the cause of an emotion via its frame annotations, it is inadequate in differentiating between different emotions, e.g. *fear* and *admire* are part of the same class. As many members of the *admire* class, e.g. *abhor*, *adore*, *deplore*, etc. – as can be seen in figure 3 – are clearly emotion-indicating verbs, we adopt many of these, labeling them with the appropriate emotion.

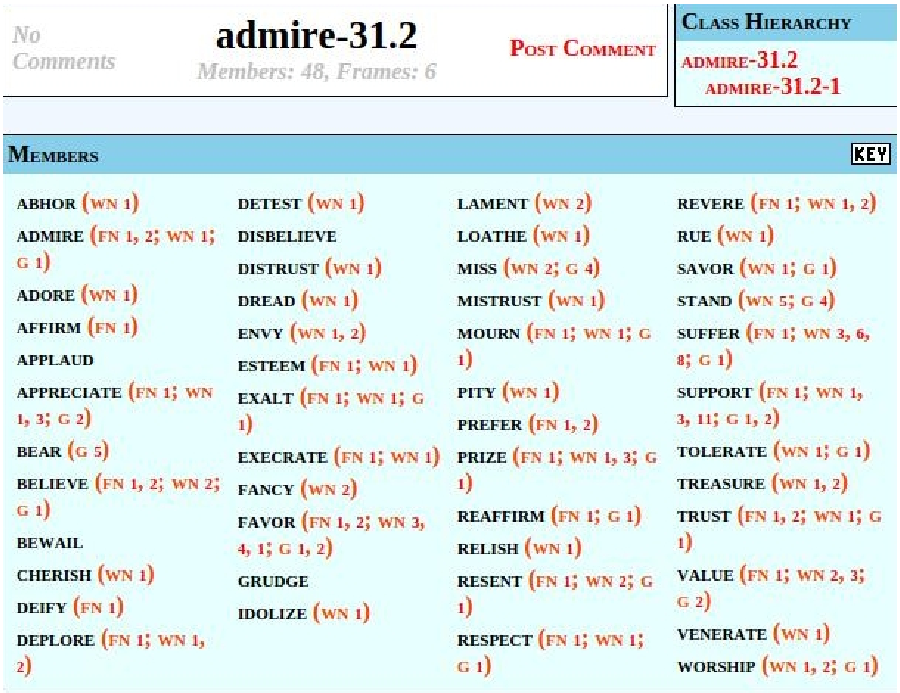

Figure 3: VerbNet's *admire* class

We also adapt many of the 253 members of VerbNet's *amuse* class, in which the subject takes the semantic role of the cause, such as *abash*, *agonize*, *appall*, etc. as these are equally evokative of emotions.

3.4.3.2 *FrameNet*

FrameNet [2] combines syntactic and semantic generalizations: Semantic frames represent the underlying meanings of the words, linking frame elements with their syntactic realizations.

In FrameNet, emotions are conceptualized in the EMOTIONS frame, which describes an Experiencer in a particular emotional State that was provoked by a Stimulus. The frame EMOTIONS is used by ten other frames. Of these, we investigate more closely the DESIRING, EMOTION_ACTIVE, EMOTION_DIRECTED, and EXPERIENCER_OBJ frames.

In the DESIRING frame, the Experiencer desires that an Event occur. We derive 20 verbs from the DESIRING frame, all of whom we label with anticipation. In the EXPERIENCER_OBJ frame, some phenomenon (the Stimulus) provokes a particular emotion in the Experiencer. Experiencer and Stimulus (cause) are core frame elements, making this frame ideal for our purposes. We derive 132 verbs from this frame. The EMOTION_ACTIVE frame is similar to EXPERIENCER_OBJ, but in this frame the verbs are more active. We derive 9 verbs from this frame.

Finally, the EMOTION_DIRECTED frame focuses on adjectives and nouns describing an Experiencer's emotional response to a Stimulus. We keep a few of these, convert some into verbs, and discard the rest, resulting in 12 verbs and adjectives from this frame.

Furthermore, EMOTIONS is inherited by the EMOTIONS_BY_STIMULUS frame that is in turn inherited by the ANNOYANCE, EMOTIONS_BY_POSSIBILITY, i.e. fear, EMOTIONS_OF_MENTAL_ACTIVITY, EMOTIONS_OF_SUCCESS_OR_FAILURE, JUST_FOUND_OUT, i.e. surprise, and OTHERS_SITUATION_AS_STIMULUS frames. While these frames all refer to emotions, we have obtained most of their lexical entries already from other sources and thus don't make use of them.

3.4.3.3 *WordNet*

WordNet [26] synsets by themselves don't contain any information about sentiment or emotion. As was previously mentioned, WordNet-Affect [47], an additional hierarchy of affective domain labels, annotates WordNet synsets representing affective concepts in the semi-automatically augmented WordNet Domains further with affective labels.

WordNet-Affect-1.1 only contains synsets that were tagged with the label EMO(TION) in the previous version and thus serve our purposes. It further labels these synsets with a set of 279 distinct emotion categories, among them uncommon emotions such as *defeatism*, *self-depreciation*, *puppy-love*, etc. These categories are hierarchically connected via `is-a` relations, which we map to Plutchik's eight emotions.

Given the ID listed in the WordNet-Affect-1.1 synsets, we retrieve the respective WordNet-3.0 synset lemmas using NLTK toolkit [5] and the corresponding category. As WordNet focuses on nouns, adjectives, verbs, and adverbs are all listed with the noun that they are derived from. For these, we thus retrieve the category of the corresponding noun. This way, we produce 784 noun-, adjective-, verb-, and adverb synsets labelled with Plutchik's eight emotions. We investigate these and make use of the verb and adjective synonyms that indicate the emotion holder as well as the cause.

SentiWordNet [13], conversely, does not prove as helpful, as we require a finer granularity than the positivity, negativity, and subjectivity scores it assigns to WordNet synsets.

3.4.4 *Summary*

We give an overview of the number of patterns we extracted from each source in table 3.

Source	# of patterns
Oxford English Dictionary	24
Merriam-Webster's Dictionary	51
Roget's Thesaurus	101
Fellbaum and Mathieu	24
Tsvetkov et al.	126
VerbNet	178
FrameNet	173
WordNet-Affect	123

Table 3: Overview of the productivity of sources for pattern design

```
fear (?! not)(?! never)scare/VB[DGPZ]/[0-9]+ NP
fear (?<= )be/VB[PDGZ]/([0-9]+)(?! not)(?! never)(
    [a-z]+/RB/[0-9]+)? scare/VBN/[0-9]+ that/IN/[0-9]+ S
fear (?<= )be/VB[PDGZ]/([0-9]+)(?! not)(?! never)(
    [a-z]+/RB/[0-9]+)? scare/VBN/[0-9]+ by/IN/[0-9]+ NP
joy  (?! not)(?! never)be/VB[DGPZ]/[0-9]+(?! not)(?! never)(
    [a-z]+/RB/[0-9]+)? happy/JJ/[0-9]+ that/IN/[0-9]+ S
```

Figure 4: Regex patterns for *fear* and *be happy about*

We remove duplicates and manually label them with an emotion, if they were not supplied with an emotion in the resource. By adding flags to indicate if the pattern refers to a nominal phrase or a clause and if the order of subject-EXPERIENCER, object-CAUSE is reversed, we transform them to the pattern template form depicted in example 5 in section 3.3.

3.5 REGULAR EXPRESSION COMPILATION

For the extraction, in order to match the patterns against our corpus described in section 3.1, we convert the pattern templates given in section 3.3 in regular expression patterns. We allow adjectives to be modified, but exclude negation. We add a regular expression capturing the index of a word to enable unambiguous retrieval of the words from the sentence. If the regular order specified by the second flag is reversed, e.g. for *annoy*, we generate additional regular expressions that capture the passive form. These are constructed with the prepositions *that* and *by* and take a clause and a nominal phrase as their complement respectively. Following this approach, we generate three regular expression patterns for *scare* and one for *be happy that*, which can be viewed in figure 4.

Generating regular expressions for both the active and passive form of patterns, if applicable, produces 662 regular expression patterns for Plutchik's eight emotions. We now match these patterns against the lemmatized, indexed, and part-of-speech tagged sentences of our corpus.

We perform a filtering of these regular expressions by running an initial matching against a random sample of 2,000,000 sentences of the corpus. As we want to guarantee accuracy as well as recall, we retain patterns with an NP cause that appear at least ten times – as these are prevalent; we retain patterns with an S cause that appear at least once.

This leaves us with 190 regular expressions, which include both active and passive forms, and 180 patterns.

3.6 MEASURING ANNOTATION AGREEMENT

We want to filter these 180 frequent patterns to keep only those that clearly refer to one emotion. To this end, we create an annotation task containing the 180 patterns that we hand off to three annotators (the author included). The guidelines of the task state that each pattern should be annotated with one of Plutchik's eight emotions. As a reference, we include Plutchik's emotion wheel depicted in figure 2. We allow the annotators to label patterns with `none` in case a pattern does not indicate an emotion. We instruct them to label the patterns with a second choice if they think that more than one emotion applies. All three annotators are instructed to label the pattern with the degree of the assigned emotion to allow for an even finer measure of agreement. Two sample entries in the annotation task can be seen in example 7, while the completed entries can be viewed in example 8.

annoy
mourn (7)

annoy anger_III
mourn sadness_I (8)

We use Fleiss' kappa (κ) as a measure of inter-annotator agreement instead of the popular Cohen's κ, as Cohen's κ only measures agreement between two annotators. Our annotators have a Fleiss' κ of 0.65, which signifies substantial agreement according to Landis and Koch. This score is particularly significant given that κ usually declines as the number of categories increases. The different levels of agreement can be observed in table 4.

As can be seen in table 5, a substantial amount, 106 of 180 expressions have been labeled unanimously with the same emotion by all three annotators. Including the second choice only increases the number slightly.

Fleiss' κ	Interpretation
0	poor agreement
0.00 - 0.20	slight agreement
0.21 - 0.40	fair agreement
0.41 - 0.60	moderate agreement
0.61 - 0.80	substantial agreement
0.81 - 1.00	almost perfect agreement

Table 4: Fleiss' κ values and their interpretations [20]

Annotated expressions	Number
Unanimous emotions	106
Unanimous emotions (including 2nd choice)	119
Majority emotions	163
Unanimous emotions + degree	39
Majority emotions + degree	131
Total	180
Fleiss' κ	0.65

Table 5: Number of annotated expressions for different forms of agreement

Quite significantly, 39 expressions have been labeled with the same emotion as well as the same degree of emotion, which is a number much higher than chance[3] for what is essentially a 25-category classification problem (8 emotions · 3 degrees + 1 (`none`)).

3.7 FINAL PATTERNS

We finally select all patterns for which we observe a majority emotion agreement, i.e. at least two of three annotators assign the same emotion. These are 163 patterns. Their distributions across all emotions can be viewed in table 6.

All of the patterns can be viewed in appendix A in table 35. Anger is relatively evenly distributed across its degrees (annoyance < anger < rage) according to Plutchik's emotion wheel (cf. figure 2); anticipation is harder to grasp, having been assigned almost exclusively to annoyance (rather than interest or vigilance) by the majority; fear appears in all its degrees, slightly less in terror; joy patterns are mostly labeled as joy, with a few having been attributed to ecstasy and only one to serenity; disgust, sadness, surprise, and trust never appear

3 For chance, only $(\frac{1}{25})^2 \cdot 180 = 0.3$ of 180 expressions would be labeled with the same degree of emotion by all annotators.

Emotion	# of majority patterns
Joy	31
Trust	8
Fear	22
Surprise	16
Sadness	18
Disgust	14
Anger	29
Anticipation	25
Total	163

Table 6: Number of patterns that have been labeled by the majority with the same emotion

in their weakest form, boredom, pensiveness, distraction, and acceptance respectively.

In general, the middle tier representing the general emotion class is prevalent among the patterns labelled by the majority, with a lot of emotions not appearing in their weakest degree due to selection bias as we tended to collect patterns that were clearly indicative of one emotion.

3.8 EXTRACTION

We extract the emotion holder and the cause using dependencies, which are well documented[4] and clearly represent the subjects, objects, and complements, which we then map to their corresponding semantic role of emotion holder or cause.[5]

We choose the Stanford collapsed dependencies for our extraction – in lieu of regular Stanford dependencies –, as these collapse dependencies like conjunctions or prepositional objects and thus facilitate extraction, e.g. a basic dependency like *prep*(cat, in) in example 9 is collapsed to *prep*(cat, hat).

I saw a cat in a hat. *prep*(cat, in)
I saw a cat in a hat. *prep_in*(cat, hat) (9)

We extract the subject via the `nsubj` (nominal subject) or `nsubjpass` (passive nominal subject) relations and the direct object via the `dobj` relation. These relations can also be embedded in disjunctions or conjunctions, which we retrieve via the collapsed `conj_and`, `conj_or`, and

4 http://nlp.stanford.edu/software/dependencies_manual.pdf
5 Initial experiments with extractions based on depth-first search in constituency trees exposed problems with noise and adaptation to sentential idiosyncrasies.

conj_but dependencies[6]. If the complement of the verb is a clause, we extract the S cause via the ccomp or xcomp relations, which are clausal complements with and without their own subjects respectively. If the verb or the adjective appears with a preposition, e.g. *take pleasure in*, *be proud of*, we retrieve the complement via the pcomp (prepositional complement) or the pobj (prepositional object) relation respectively. If the verb modifies the subject, e.g. *a ruler loved by the people*, we extract the subject via the partmod relation[7].

Besides the emotion holder and cause, we want to extract all relevant information that might be useful for further analysis. Borth et al. clearly show that adjectives have significant emotive value by pairing positive or negative adjectives with neutral nouns to construct their adjective-noun pairs as described in section 2.1. Thus we capture all modifiers as well as prepositional objects using the following relations:

A. the modifiers of the noun:
   a) nn (noun compound modifier), e.g. nn(University-10, Harvard-9),
   b) amod (adjectival modifier),
   c) num (numeric modifier)[8];

B. the prepositional objects pertaining to the noun and the verb; collapsed prepositional objects consist of the preposition name prefixed with prep_, e.g. prep_on.

We insert adjectival and numeric modifiers before the noun and append prepositional objects prefixed by the preposition and separated by an underscore, if it is a noun, e.g. for:millionth_fan. We include prepositional objects modifying the predicate of a clause in a separate bag-of-words.

Often, the emotion holder and the cause of an emotion are expressed by pronouns, which by themselves only provide limited information and generalizability. For this reason, we leverage the Stanford coreference resolution annotations of the Annotated Gigaword replacing every mention with its most representative coreferent.

To further make our results more generalizable, we tag every named entity (NE) with its NE tag, i.e. LOCATION, PERSON, and ORGANIZATION, and replace numbers with the NUMBER tag.

Dependent on the argument order, we assign the subject and the object/clause to the EXPERIENCER or CAUSE respectively; we discard extractions where the cause still contains a pronoun, as we require our extractions to be as generalizable as possible.

6 The uncollapsed cc relation would increase the search depth by another relation.
7 As of the Stanford dependencies version 3.5.2, partmod has been generalized as a case of vmod (reduced non-finite verbal modifier)
8 num is converted to nummod as of version 3.5.2.

```
h5n1 strain trigger global influenza pandemic
[the/DT, h5n1/JJ, strain/NN, will/MD, trigger/VB, a/DT,
    global/JJ, influenza/NN, pandemic/NN, that/WDT, could/MD,
    kill/VB, million/NNS, around/IN, the/DT, world/NN]
```

Figure 5: Extracted cause and bag-of-words of the cause from an example sentence

```
NYT_ENG_19960601.0010/1 trust rely on Japan/LOCATION economic
    clout     [its/PRP$, economic/JJ, clout/NN]
NYT_ENG_19960601.0010/9 anticipation expect Japanese    win
    right  [to/TO, win/VB, the/DT, right/NN, to/TO, host/VB,
    the/DT, tournament/NN]
```

Figure 6: Extractions with NP and S cause

Finally, we retrieve all leaves of the phrase-structure tree of the sentence that are dominated by the node of the cause (including subordinate clauses) as well as the node of the cause itself and list them in a bag-of-words tagged with their parts-of-speech. The cause and the bag-of-words representation of the cause for the following sentence are depicted in example 5:

World Health Organization experts fear the H5N1 strain will trigger a global influenza pandemic that could kill millions around the world. (`APW_ENG_20050303.0369/8`).

While the cause focuses on the main arguments of the complement, the bag-of-words includes the subordinate clause *that could kill millions around the world*, which provides additional information regarding the source of the emotion.

3.9 REPRESENTATION OF EXTRACTIONS

Using this pattern-based extraction approach, we obtain from the following sentences the extractions listed in figure 6 respectively (with tabs as delimiters):

- *The countries had engaged in a multimillion-dollar battle to host the tournament, with Japan relying on its economic clout and South Korea relying on its superior soccer pedigree.*
- *The Japanese, who expected to win the right to host the tournament, were dismayed.*

In both examples, Japan is the emotion holder. In the first instance, the cause is a nominal phrase extracted via the `dobj` relation, while in the second one, the cause is a clause extracted using the `xcomp` relation. We provide the following properties in ascending order of column index:

a. a unique identifier for each extraction created by concatenating the Annotated Gigaword document ID with the index of the sentence in that document;

b. the emotion;

c. the pattern that produced the extraction;

d. the emotion holder;

e. the NP cause of the emotion (can be empty);

f. the subject of the S cause of the emotion (can be empty);

g. the predicate of the S cause of the emotion (can be empty);

h. the direct object of the S cause of emotion (can be empty);

i. prepositional objects modifying the predicate of the S cause (can be empty); and

j. a bag-of-words representation of the cause, tagged with parts-of-speech.

As we also collect the full sentences, we can easily cross-reference extractions via the ID to make use of parts of the proposition that have not been extracted initially.

4 ANALYSIS

We have described in chapter 3 the process of laying the seeds for our work by designing and evaluating the patterns for our extraction. In this chapter, we are now finally able to reap the rewards and harvest the emotion propositions.

For this purpose, we will perform different distributional analyses of the extracted propositions regarding a variety of aspects to gain further insights. In section 4.1, we will introduce our resource, the Emotion Proposition Store, and give an overview of the nature and distribution of the extracted propositions as well as evaluate the productivity of the patterns that produced them. Subsequently, in section 4.2, we will compare two association metrics frequently employed in sentiment analysis, point-wise mutual information (PMI) and chi-square, for their use in determining ngrams that are most associated with an emotion and evaluate the capability of unigrams versus bigrams to evoke emotions. We store these ngrams in lists that we make available to the research community, which can be used as an emotion lexicon for the news domain. In section 4.3, we evaluate the prior probability of association with an emotion for the following configurations of the emotion holder and cause: (a) emotion holder; (b) NP cause; (c) S cause subject + predicate; and (d) S cause predicate + object. To this end, we use two evaluation methods: (a) We match our bigrams against annotations in the emotion lexicon EmoLex [30] that we presented in section 3.4.2.2 to obtain an initial guidance. (b) Secondly, we let human annotators annotate the most promising configurations of bigrams to obtain a gold standard that we use for evaluation. Finally, in section 4.4, we use an extension of Latent Dirichlet Allocation (LDA) that allows the introduction of supervised category information and enables us to generate underlying topics that are associated with our emotions. We investigate different topic configurations and intrinsically evaluate these topics using EmoLex.

4.1 EXTRACTION RESULTS

Matching the patterns against the whole Annotated Gigaword corpus produces 2,320,636 propositions. These extracted propositions, however, contain many duplicates, e.g. the sentence *In 1997, with only three years as Texas governor, Bush raised a record-sized warchest that scared away several challengers and found himself at the top of polls of potential 2000 GOP candidates* occurs 25 times in the annotated Gigaword file APW_ENG_200011.XML.GZ in different documents.

A removal of all extractions that stem from identical sentences yields 1,788,022 extractions. After an investigation of the results, we decide to remove extractions with the following patterns from the corpus as they may often refer to unemotive contexts in the news domain that are different from the ones we originally intended:

- *grate*: Although occasionally used emotively, *grate* is most often used in an unemotive context related to cooking.
- *depress*: As economics are a frequent topic, *depress* often refers to sales or prices rather than the mental state of persons.
- *aggravate*: In international politics, *aggravate* often relates to conditions instead of individuals.
- *rattle*: In politics someone may *rattle off* an argument, or an earthquake may *rattle* a city.
- *afflict*: An ailment often *afflicts* people.
- *inflame*: *Inflame* is often used metaphorically, i.e. *inflame* the bonfire of terrorism, the spirit of war, etc.

Emotion	**Frequency**	**% of total extractions**	**# of patterns with 10+ occurrences**	**total # of patterns**
anticipation	966,571	54.47	22	25
fear	249,103	14.04	20	22
joy	231,967	13.07	30	31
trust	89,217	5.03	6	8
anger	64,586	3.64	28	29
surprise	60,221	3.39	20	16
disgust	59,486	3.35	14	14
sadness	53,269	3.00	13	18
Total	1,774,420	100.00	153	163

Table 7: Frequencies of emotions in extractions

Excluding these patterns in their active and passive forms leaves us with 1,774,420 distinct propositions. These propositions form our Emotion Proposition Store, which we make available to the research community.[1] We also release the entirety of the extracted sentences to

1 It can be found at https://github.com/sebastianruder/emotion_proposition_store/tree/master/out/emotion_proposition_store in *shelves* of 100,000 propositions.

enable the contextualization of propositions.[2] The emotion distribution of the propositions is depicted in table 7. As can be seen, anticipation makes up slightly more than half of all extractions, which is mainly due to *expect* being a very productive pattern, as depicted in table 9. Fear and joy both occur frequently in propositions, while sadness, disgust, and surprise are less pronounced. The patterns follow a Zipf distribution, with the two most frequent patterns generally accounting for more extractions than the remaining patterns combined. Significantly, even though we gather the fewest propositions for sadness, it achieves the most accurate results with our association metrics as displayed in table 11.

As we intend to investigate emotion association for concepts as well as events, we have clearly distinguished between nominal phrases and clauses as causes. We show their distribution among the extracted propositions in table 8.

Emotion	**# extractions with NP cause**	**# extractions with S cause**
anticipation	407,738	558,833
joy	190,484	41,483
fear	82,116	166,987
trust	72,483	16,734
surprise	59,657	564
disgust	58,942	544
anger	57,379	7,207
sadness	26,064	27,205
Total	956,392	819,557

Table 8: Patterns with NP vs. S cause

While anticipation prevails for both nominal and clausal causes, joy propositions are skewed towards nominal phrases, while fear propositions are dominated by clauses. We find that, except for anticipation and fear, nominal causes dominate – and quite clearly in the cases of surprise, disgust, and anger. Most of the patterns of these emotions consist of Experiencer psych verbs such as *surprise*, *anger*, *provoke*, etc., which take the emotion holder, i.e. a nominal phrase, as their argument. Style conventions of news texts and creative writing discourage the use of passive forms, which are more apt in introducing clauses. Consequently, these are not found in the top ten patterns of any emotion, except for trust, which had fewer than ten patterns. Furthermore, nominalization or the conversion of clauses to appro-

2 They can be accessed at https://github.com/sebastianruder/emotion_proposition_store/tree/master/out/sentences. Note that duplicate sentences have not been removed as they are meant for cross-reference.

priate nominal phrases is a common feature of news texts, i.e. in *I hate the idea of seeing (Pluto) expelled from the society of certified planets* (NYT_ENG_19960311.0669/31), the object of hatred is a nominal phrase rather than the clause *seeing it expelled from the society of certified planets*.

In the following, we display the top 10 patterns for each of Plutchik's eight emotions and discuss them.

Pattern	Frequency
expect	490,640
predict	156,681
await	71,910
look forward to	59,884
prepare for	45,687
forecast	45,365
hope	28,277
anticipate	19,112
be hopeful	17,732
foresee	9,248
Total top 10	944,536
Total	966,571

Table 9: Top 10 anticipation patterns

Pattern	Frequency
surprise	18,358
stun	16,211
shock	13,269
confuse	2,899
wow	2,003
confound	1,771
startle	1,680
baffle	1,169
puzzle	802
stagger	689
Total top 10	58,851
Total	60,221

Table 10: Top 10 surprise patterns

Analogous to social media data, which has been found to be predominantly happy (cf. [45]) due to its publicity that incentives the sharing of positive messages for gratification, neutral predicates are prevalent given the idiosyncrasies of the objective news domain. In fact, journalistic objectivity encompassing qualities such as fairness, factuality, and nonpartisanship, forms a significant principle of journalistic professionalism. Consequently, anticipation as well as neutral predicates such as *expect*, *predict*, and *rely on* rank highly.

Surprise, the second neutral emotion, occurs significantly less often than its counterpart, anticipation, as the overt expression of surprise is typically linked to unpreparedness, which can be associated with weakness or incompetence, properties generally undesired to be exposed in public. The most prominent surprise patterns, the eponymous *surprise*, as well as *stun*, and *shock* occur significantly more often than the remaining ones.

Joy and fear, the two second most frequent emotions, exhibit the same particularity, in that the most frequent pattern occurs more often than all of the others combined. *Enjoy* and *fear* thus seem to be the preferred ways to indicate the polar opposites of joy and fear, as they make the emotional reaction explicit.

Pattern	Freq
enjoy	131,102
be proud	21,842
be lucky	13,857
cheer	10,658
relish	5,590
impress	5,431
entertain	4,318
arouse	4,247
be satisfied	4,016
please	3,394
Total top 10	204,455
Total	231,967

Table 11: Top 10 joy patterns

Pattern	Frequency
fear	155,968
be afraid	35,314
worry	25,361
be anxious	10,427
scare	6,704
be scared	3,340
be nervous	2,195
frighten	1,992
unnerve	1,573
dread	1,534
Total top 10	244,408
Total	249,103

Table 12: Top 10 fear patterns

Pattern	Freq
rely on	50,480
count on	17,562
trust	14,749
reassure	5,384
take comfort in	620
charm	412
be charmed	7
be reassured	3
Total top 10	89,217
Total	89,217

Table 13: Top 10 trust patterns

Pattern	Freq
hate	22,804
shun	9,268
deplore	8,751
dislike	6,069
alienate	2,675
repel	2,054
despise	1,828
loathe	1,328
disdain	1,208
abhor	1,149
Total top 10	57,134
Total	59,486

Table 14: Top 10 disgust patterns

Trust patterns, even though we were only able to identify eight frequent and unambigous ones, occur less often than Plutchik's emotional opposite (cf. figure 2), disgust. *Rely on* is a particularly productive pattern, that surpasses *count on* and *trust*. The most frequent disgust pattern in turn, *hate*, does not occur significantly more often than the rest. Other frequent disgust patterns like *shun*, *deplore*, and *dislike* introduce variance in the expression of disgust. In contrast to fear, disgust is not an emotion that is frequently expressed in news.

Pattern	Freq
regret	24,395
mourn	7,948
be unhappy	6,120
be sad	5,071
disappoint	3,320
upset	2,439
grieve	1,317
pine for	1,064
agonize over	489
displease	423
Total top 10	52,586
Total	53,269

Table 15: Top 10 sadness patterns

Pattern	Frequency
provoke	16,528
be angry	6,808
bother	6,612
infuriate	5,316
harass	4,791
insult	3,534
frustrate	2,912
irritate	2,805
offend	2,482
fret	2,356
Total top 10	54,144
Total	64,586

Table 16: Top 10 anger patterns

Finally, the feeling of remorse or regret is the most common manifestation of sadness in news, while grief, expressed through *mourn* is felt to a lesser degree. The prominence of the anger pattern *provoke*, which frequently brings about a reaction, conforms to the reporting style of news.

4.2 DISTRIBUTIONAL ANALYSIS OF CAUSES

We will now divert our attention from the patterns to the propositions that we extracted, particularly the causes as these are most prone to be associated with an emotion. For this purpose, we will perform various distributional analyses in this section. We will use two association metrics commonly used in sentiment analysis to calculate the degree of association between an emotion and ngrams derived from the nominal or clausal causes. Ordering the ngrams by their association score will allow us to determine the concepts, actions, and events that are most associated with certain emotions. A comparison of high-ranked unigrams and bigrams will shed light on their propensity to be associated with an emotion. By aggregating the association score across all emotions, we will reveal ambiguous and controversial concepts.

We derive unigrams from the sources depicted in table 17. These are, for unigrams, the emotion holder, the NP cause, and the S cause. We keep these for bigrams and add bigrams consisting of the subject of the S cause concatenated with its predicate, and the predicate of the S cause concatenated with its object. In both of the latter, the predicate is always part of the bigram, as it is the most meaningful part in a clausal cause.

Source	Unigram	Bigram
Emotion holder	X	X
NP cause	X	X
S cause	X	X
S cause subject + predicate		X
S cause predicate + object		X

Table 17: Sources for unigram and bigram generation

We count the occurrences of all such unigrams and bigrams. In order to strike a balance between reliable results and context-sensitivity[3], we treat prepositional objects as one unigram, e.g. *of:innocent_life* is treated as one unit, instead of three separate unigrams.[4]

We convert all ngrams to lowercase if they are not named entities; for named entities, we keep the NE tags in a shortened form for better readability[5]; finally, we remove all stopwords.

The counts of number of unigrams and bigrams for the different configurations can be found in table 18. We observe unigram counts around 2,000,000, while bigram counts range around 1,000,000, except for the predicate + direct object configuration.

Ngram type	# unigrams	# bigrams
Subj + Pred (S cause)	0	921,337
NP cause	1,856,712	915,122
Pred + Dobj (S cause)	0	168,013
S cause	2,307,214	1,490,232
Emotion holder	2,809,796	1,279,699

Table 18: Number of unigrams and bigrams for different emotion holder / cause configurations

In order to determine the concepts, actions, and events that are most associated with certain emotions, we use two association metrics that have been used frequently in sentiment analysis: pointwise mutual information (PMI) and chi-square.

3 Initial experiments revealed that we lack sufficient data to obtain reliable results for trigrams, which would provide additional context helpful for emotion detection.

4 Another reason for choosing this representation is that bigrams such as *loss of:innocent_life* are significantly more relevant than its alternative variations, *loss of* or *innocent life*.

5 i.e. `ORGANIZATION` → `ORG`; `PERSON` → `PERS`; `LOCATION` → `LOC`; `NUMBER` → `NUM`.

4.2.1 *PMI*

PMI, which is derived from information theory, models the mutual information between two expressions. PMI for two expressions x and y is defined as the ratio of their joint probability and the product of their individual probabilities:

$$pmi(x;y) = log\frac{p(x,y)}{p(x)p(y)}$$

with $p(x,y) = \frac{f(x;y)}{N}$ and $p(x)p(y) = \frac{f(x)f(y)}{N^2}$, where $f(x)$ is the frequency of x in the corpus and N is the number of extractions.

x and y are associated with each other, if their PMI is greater than 0 – with their association strength increasing with the PMI score; they are negatively associated if their PMI is less than 0. For instance, sadness would be negatively associated with a concept that is prone to evoke joy, e.g. ice cream. In our case, x is the emotion and y is a unigram or bigram that occurs with the emotion.

The regular PMI measure has one issue in that if y occurs just once with emotion x and with no other emotion, $p(x,y)$ equals $p(y)$. This reduces the above equation to $pmi(x;y) = log\frac{1}{p(x)}$, which is independent of the ngram y as it is just the probability of emotion x.

Mohammad simply ignore ngrams that occur less than five times, which, however, only unsatisfactorily mitigates the issue. Instead, in order to penalize such low-frequency observations, we use a simple discounting method that has been described by Pantel and Ravichandran and already put to you use by Gerber and Chai. Multiplying the numerator with the discount factor produces the following term:

$$\frac{f(x;y)}{N} \cdot \frac{f(x;y)}{f(x;y)+1}$$

If the ngram occurs very rarely, the discount factor will be small, reducing the PMI score. We apply a similar discount to the denominator, which reduces to the n-gram frequency divided by the n-gram frequency + 1, as the ngram would always occur less often than the associated emotion category:

$$\frac{f(x)f(y)}{N^2 \frac{min(f(x),f(y))}{min(f(x),f(y)+1})} = \frac{f(x)f(y)}{N^2 \frac{f(y)}{f(y)+1}}$$

If the ngram is rarely observed, $\frac{f(y)}{f(y)+1}$ is small, making the denominator large, which reduces the PMI score. In general, the discount factor reduces the PMI score for ngrams that are very infrequent ($f(y)$ close to 1) and minimally impacts all other PMI scores – which is exactly the behaviour that we want.

Another common variant of the PMI measure is the normalized PMI score, which restricts the PMI value to the interval $[-1,+1]$, where -1 signifies no joint occurrence, 0 is independence, and +1 represents complete co-occurrence. We have experimentally applied this, but found that it negates the effect of the introduced discount, which led us to decide against its use.

4.2.2 *Chi-square*

Chi-square (χ^2) is another common measure to gauge the association between two concepts. Let n be the number of all documents, i.e. extractions, $p(x|y)$ be the conditional probability of emotion x occurring with ngram y, $P(x)$ be the fraction of extractions containing emotion x, and $F(y)$ be the fraction of extractions containing ngram y, which we define as follows:

$$p(x|y) = \frac{p(x;y)}{p(y)}$$

$$P(x) = \frac{f(x)}{n}$$

$$F(y) = \frac{f(y)}{n}$$

With these, the χ^2 metric between emotion x and ngram y is defined as [1]:

$$\chi^2 = \frac{n \cdot F(y)^2 \cdot (p(x|y) - P(x))^2}{F(y) \cdot (1 - F(y)) \cdot P(x) \cdot (1 - P(x))}$$

4.2.3 *PMI vs. chi-square*

χ^2 and PMI are two different ways of measuring how well a term and a category correlate. χ^2 in contrast to PMI is a normalized value, making it more comparable across the same category. χ^2, however, doesn't penalize infrequent terms, which the discounted PMI score does.

We calculate PMI and chi-square scores for unigrams and bigrams derived from parts of the proposition depicted in table 17. We order these by metric, emotion, source, and ngram and rank them by their association score. We store them in emotion-labelled lists that we make publicly available.[6] These lists can be used as an emotion lexicon that is particularly apt to gauge emotion association in the news domain. Their finely granular categorization moreover makes

6 They can be accessed at https://github.com/sebastianruder/emotion_proposition_store/tree/master/out/scores. Note that the third and fourth column of the files contain emotion and sentiment overlap values relevant for the evaluation described in section 4.4.3.

them ideal for identifying certain semantic roles in emotion detection, while the bigrams provide increased contextualization versus the unigrams contained in other emotion lexicons. We furthermore generate charts that show the distribution of the scores and for the highest-ranked ngrams, which we release likewise.

We compare the scores of the PMI and χ^2 metrics for an initial set of ngrams and then proceed with the metric that yields the better performance. We clearly want to be able to compare values not only among the different types of unigrams and bigrams that we extract but also across different emotions. If we compare the PMI values of top bigrams in the NP cause of different emotions, we find that their range lends itself to easy comparison, with the top 10,000 bigrams of each emotion falling in an interval between 3.40 (sadness) and 0.16 (surprise) as depicted in figure 7.

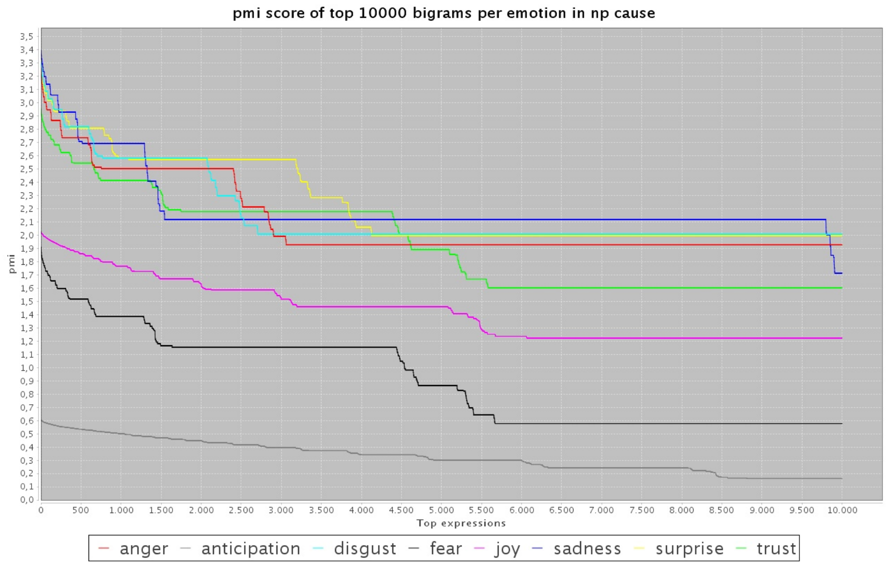

Figure 7: PMI values for the top 10,000 bigrams for each emotion

Sadness, surprise, disgust, anger, and trust all show a similar slope with top bigrams ranging from 3.40 (sadness) to 2.95 (trust) and then dropping off steadily. Joy bigrams start at 2.03, while fear bigrams begin at 1.91 and anticipation bigrams start at a low point of 0.60. However, high overall PMI scores for an emotion are not necessarily correlated with increased accuracy, as we achieve good results for both joy and fear as depicted in table 30. The flat curve of anticipation, however, shows that concepts are only weakly associated with it.

Conversely, the distribution of χ^2 bigrams for each emotion has an extremely long tail that converges towards 0 and no substantial score differences can be observed after a threshold of 100. We thus show only the top 100 bigrams with the highest χ^2 values for each emotion. The distribution of these is extreme, ranging from 9,426 (sadness) to 73.68 (fear), as can be seen in figure 8.

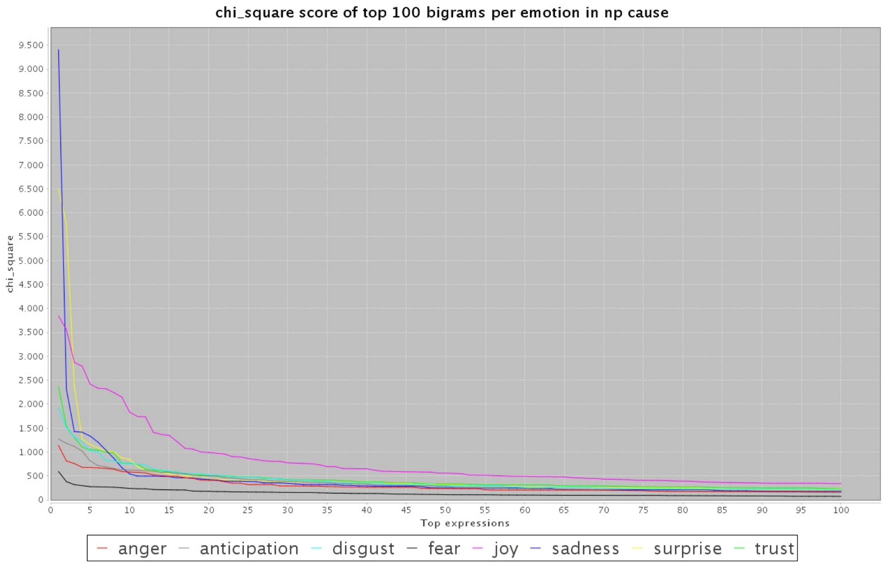

Figure 8: χ^2 values for the top 100 bigrams

High-ranked sadness and surprise bigrams have the highest χ^2 values with 9,426 and 6,496 respectively, but drop sharply below 1,000 to the level of the other emotions after the top 10 bigrams, when the slope flattens. Joy bigrams start off third-highest at 3,842 and decline less rapidly, adapting to the others only after the top 100 bigrams. Even as χ^2 values start to become more similar in lower ranks across emotions, the highest-ranked bigrams are outliers that skew the data, particularly as we would like to aggregate χ^2 scores of bigrams, essentially rendering the comparison of χ^2 values across emotions moot.

After gauging the distribution of the association scores across the highest-ranked bigrams, we now inspect the actual bigrams to gain a better understanding of their quality. We choose to compare bigrams for sadness, as it achieved the top-ranked bigram for each measure. We list the top 10 NP cause bigrams of sadness produced by χ^2 and PMI in table 19.

Both metrics produce sensible candidates. Bigrams such as *death of:NUM* and *death of:NUM_people* appear among the highest-ranked terms for both association metrics, highlighting a comparable performance. Furthermore, both contain bigrams dealing with *loss* and *victims*. Among χ^2 sadness bigrams, the theme of *loss* is prevalent, while among PMI bigrams *death* dominates. PMI bigrams are also grouped much more closely, while the huge gap between the first and second χ^2 candidate is not plausibly justifiable. If there was a significant difference at all, *loss of:innocent_life* should be more associated with sadness than *loss of:life*. In light of these issues, particularly the sharp decline of χ^2 scores for the same emotion and the resulting incomparability across emotions, we choose to proceed with further analyses using the PMI metric.

χ^2 **bigram**	χ^2 **value**	**PMI bigram**	**PMI value**
loss of:life	9426.09	death of:NUM	3.40
loss of:innocent_life	2317.45	quake victim	3.37
tragic loss	1430.38	death of:wife	3.37
loss of:civilian_life	1416.05	loss of:man	3.37
civilian casualty	1303.09	tragic loss	3.36
death of:NUM	1208.49	death of:father	3.36
death of:NUM_people	1038.87	death of:NUM_people	3.34
NUM victim	870.43	death of:relative	3.33
unfortunate incident	677.81	loss of:friend	3.33
choice of:word	516.71	innocent victim	3.33

Table 19: Comparison of χ^2 and PMI values for the top 10 sadness NP cause bigrams

4.2.4 *Bigrams vs. unigrams*

The bigrams depicted in table 19 are clearly associated with sadness. Emotion lexicons such as the ones mentioned in section 2.6, however, only contain unigrams. We thus inspect the corresponding unigrams, the top 10 NP cause sadness unigrams, in table 20 to gain an understanding of their suitability to convey emotions.

Unigram	**PMI score**
inconvenience	3.43
misunderstanding	3.42
passing	3.32
to:embassy	3.32
slay	3.28
of:grandson	3.27
assassinate	3.27
of:Princess_Diana/PERS	3.25
error	3.25
uday	3.24

Table 20: Top 10 NP cause sadness unigrams

While unigrams such as *passing* and *slay* are clearly associated with sadness, for others this relation is harder to fathom without context. Bigrams like *error*, *inconvenience*, and *misunderstanding* could likewise be the source of different degrees of anger or disgust rather than sadness. This is even more pronounced for others such as *to:embassy*, *of:grandson*, or *of:Princess_Diana/PERS*, which are inconsequential without the relevant context. Note that *uday* – actually Uday Hussein – was erroneously tagged as VBP by the Stanford parser.

In order to obtain more meaningful results in further analyses, we focus on bigrams instead of unigrams, as their additional context improves their aptitude of being associated with an emotion.

4.2.5 *Ambiguous concepts*

Before investigating individual emotions, we take a look at the most ambiguous bigrams, those which have the highest PMI score summed across all emotions. The top 20 NP cause bigrams are displayed in table 21, with each emotion's portion of the overall PMI score depicted in percent. The same information is displayed in absolute PMI scores for the top 50 bigrams in figures 9 and 10 across emotions and sentiment respectively. Interestingly, anticipation and fear don't appear among the most ambiguous bigrams. We will elaborate on this below.

Bigram	**anger**	**disgust**	**joy**	**sadness**	**surprise**	**trust**
former president	21.09	28.12	0.79	20.74	25.46	3.81
prime minister	28.50	11.05	0.00	2.71	31.20	26.53
shoot death	27.55	20.80	0.00	31.05	20.60	0.00
United/LOC States/LOC	19.35	34.57	0.00	0.00	25.77	20.31
hostage crisis	29.78	22.87	0.00	24.69	22.66	0.00
president Chen/PERS	41.25	32.82	0.00	0.00	19.22	6.72
president Clinton/PERS	27.69	32.52	1.43	11.26	13.96	13.16
Tony/PERS Blair/PERS	21.51	28.01	0.00	0.00	25.61	24.87
police officer	34.97	18.29	0.00	2.19	34.56	9.99
president George/PERS	21.21	25.61	1.68	0.00	26.18	25.32
Israeli/LOC prime	30.29	0.00	0.00	12.08	37.24	20.39
Boris/PERS Yeltsin/PERS	0.00	19.69	0.00	13.78	43.14	23.39
security guard	16.49	42.61	0.00	0.00	18.10	22.79
russian troops	14.11	33.98	0.00	0.00	26.94	24.98
minister Tony/PERS	25.29	23.48	0.00	0.00	23.27	27.96
president Jacques/PERS	36.64	15.50	0.00	21.97	9.19	16.69
Rudolph/PERS Giuliani/PERS	30.31	32.14	0.00	0.00	37.54	0.00
W./PERS Bush/PERS	14.72	30.69	4.48	1.38	25.89	22.84
Saudi/LOC Arabia/LOC	25.35	0.00	0.00	9.31	38.51	26.83

Table 21: Emotion distribution in % of top 20 NP cause bigrams with highest aggregated PMI score; *fear* and *anticipation* have 0 PMI for all instances

Named entities like state leaders and states are among the most polarizing concepts. Even regarding these political powers, PMI scores of the neutral anticipation – essentially the most common emotion in politics – are negative, emphasizing the genericity and vagueness of Plutchik's anticipation and underlining why it isn't used in other classifications, e.g. Ekman's [11]. Even the strongest examples of association are only comparatively weakly associated with it, as depicted in 7.

If we ignore named entities, expressions that are associated with violence such as *shooting death* and *hostage crisis* or that may potentially be deployed controversially such as *police officer*, *security guard*, and *Russian troops* receive the highest aggregated scores. Trust is placed in the *Russian troops* e.g. by *Tajikistan 's hard-line government*, which relied on them in 1998 *to guard the Afghan border against incursions by guerillas and smugglers* (`APW_ENG_19980116.0782/13`). The capability to showcase the development of such emotive trends over time thus presents an interesting research opportunity and will produce further insights on how public opinion on certain concepts fares. Date information can be easily extracted from the extraction ID; as an investigation of the temporal aspect would significantly inflate the scope of this project, we will leave it for future work.

Interestingly, even among such potentially violent concepts as the ones mentioned above, PMI scores for fear are still negative. Conversely, the top 10 NP cause bigrams that have the highest fear PMI scores (cf. appendix A, table 36) all spread fear on a national level, e.g. *government reprisal*, *political backlash*, *reprisal attack*, etc. while the above mentioned ambiguous concepts are condemned by the general public rather than scare it. An explanation for this may be that fear mainly channels other negative emotions such as anger and disgust, as its intense nature prevents it from being strongly associated with other positive or neutral emotions, which would be required in order to occur among the top controversial concepts. Indeed, *missile strike* and *suicide bomber*, two of the first bigrams that show moderate degrees of fear also exhibit anger and disgust and – surprisingly – the positive trust. Closer inspection reveals that the U.S. relies on missile strikes, while foreign-led organizations such as the Hezbollah depend on suicide bombers. Two other terms, *Liberation Army* and *rocket attack* evoke fear as well as anger, disgust, and surprise, an emotion that – due to its neutrality – can have both a positive or negative connotation dependent on the context.

Finally, considering the two expressions ranked as the most ambiguous, *former president* and *prime minister*, both of them can refer to leader of states that are objects of a diverse set of emotions, e.g. *South Korea's former president Roh* (`AFP_ENG_20090528.0118/7`) is mourned by thousands, while Washington loathes *two (corrupt) former presidents (of Nicaragua)* (`APW_ENG_20051205.0005/5`). Even former leader's deci-

Emotion	Sentiment	Emotion	Sentiment
anticipation	neutral	anger	negative
surprise	neutral	fear	negative
joy	positive	disgust	negative
trust	positive	sadness	negative

Table 22: Mapping of emotions to sentiment

sions may still evoke expressions of anger or surprise, but the media will be more cautious with signs of the stronger disgust for current than for former leaders. Furthermore, the degree of anteriority implied by *former president* is linked to sadness, whereas support will rather be given to a current leader.

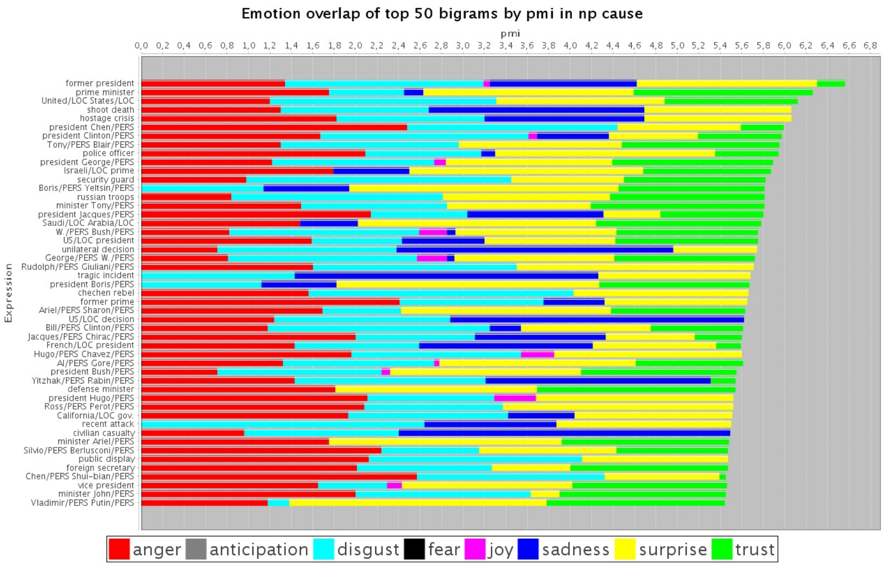

Figure 9: Overlap across emotions for the top 50 PMI NP cause bigrams

Considering the entirety of ambiguous expressions displayed in figure 9, we observe that degrees of anger, disgust, and surprise are present in almost all concepts. Trust frequently appears as well, sadness less often, although more dominantly if it does, while joy is hardly visible. In order to provide a higher-level distribution of these concepts, we map Plutchik's eight emotions to their corresponding sentiment in table 22.

We show the distribution of sentiment across the ambiguous concepts in figure 10.

As we can see, negative sentiment prevails. Neutral sentiment is associated with all but three concepts, while positive sentiment only occurs in a slight majority.

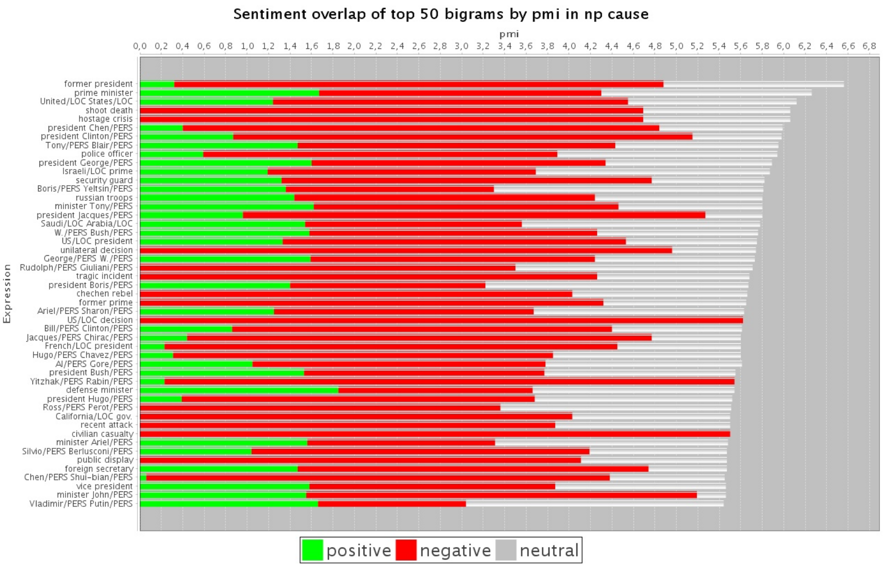

Figure 10: Overlap across sentiment for the top 50 PMI NP cause bigrams

4.3 EVALUATION

Having compared different association metrics as well as unigrams and bigrams, we will evaluate the quality of the highest-scoring bigrams for each emotion. For this purpose, we use two evaluations:

A. We will evaluate against the NRC Word-Emotion Association Lexicon (EmoLex) [30] that we introduced in section 3.4.2.2. As EmoLex lists unigrams together with the emotion and sentiment that they are associated with, it serves as a good starting point for our evaluation. As our lists contain the same information, we can perform an initial evaluation via a simple look-up: We check if one of the unigrams in our bigrams appears in EmoLex and if so, we check if they are associated there with the same emotion and corresponding sentiment that we assigned. As bigrams' emotion often is compositional, this evaluation mostly serves to inform which bigram sources are most promising rather than to exactly ascertain their accuracy.

B. We will review these bigrams in a second step using a more rigorous evaluation that employs manually annotated data. Annotators will perform an annotation task to label bigrams with an emotion; an evaluation against this manually-created gold standard will then provide a more objective overview of the quality of our bigrams.

.

Bigram	Emotion overlap	Sentiment overlap
death of:number	TRUE	TRUE
coup today	FALSE	FALSE
arrogant response	FALSE	TRUE

Table 23: Overlap with EmoLex for sample sadness bigrams

4.3.1 *Evaluation using EmoLex*

Regarding the evaluation using EmoLex, we evaluate bigrams from the sources depicted in table 17, (a) the emotion holder, (b) the NP cause, (c) the S cause subject and predicate, and (d) the S cause predicate and object. We exclude bigrams that contain a named entity as these are not present in EmoLex. In order to facilitate a subsequent investigation of the results, we select the top 30 bigrams per emotion per above mentioned source and calculate a ternary overlap with EmoLex: As EmoLex only contains unigrams, we assign `NA`, if none of the unigrams of the bigram appears in EmoLex; if one of them appears in EmoLex with the respective emotion, we assign `TRUE`; otherwise we assign `FALSE`. As EmoLex also contains a label for *positive* and *negative*, we also calculate the sentiment overlap by checking if the emotion's corresponding sentiment (see table 22) appears with the respective sentiment in EmoLex as above.

To illustrate this procedure, we show the overlap with EmoLex for three sadness bigrams in table 23. *Death* in *death of:number* appears in EmoLex and is labelled with sadness as well as a negative sentiment and thus is assigned `TRUE` for both emotion and sentiment overlap. Both *coup* and *today* appear in EmoLex, but neither of them is assigned sadness nor a negative sentiment; for this reason, *coup today* is assigned `FALSE` for both emotion and sentiment overlap. Finally, *arrogant response* receives a negative emotion overlap, but a positive sentiment overlap, as neither *arrogant* nor *response* are associated with sadness in EmoLex, while *arrogant* is associated with a negative sentiment.

We show the results in tables 24, 25, 26, and 27. We include figure 11 exemplarily to visualize the values in table 27. Sentiment columns for anticipation and surprise only contain `NA` as EmoLex does not contain a neutral label. All numbers are given as percentage of the top 30 bigrams that were chosen for each emotion respectively. Note that percentages for `TRUE` equal precision. As our gold standard for this evaluation, EmoLex, contains 14,000 terms and we only select 30 bigrams per emotion for this evaluation, we do not provide recall nor F-score as they would not add additional insights.

As we have outlined before, emotion is a compositional phenomenon. Thus, this evaluation employing comparison on an unigram-level only serves as a guideline that will inform further investigation.

Overlap	**anger**		**anticipation**		**disgust**		**fear**	
	Emo	*Sen*	*Emo*	*Sen*	*Emo*	*Sen*	*Emo*	*Sen*
TRUE	46.67	50.00	10.00	0.00	6.67	53.33	10.00	20.00
NA	6.67	6.67	50.00	100.00	20.00	20.00	20.00	20.00
FALSE	46.67	43.33	40.00	0.00	73.33	26.67	70.00	60.00
Overlap	**joy**		**sadness**		**surprise**		**trust**	
	Emo	*Sen*	*Emo*	*Sen*	*Emo*	*Sen*	*Emo*	*Sen*
TRUE	3.33	30.00	3.33	6.67	0.00	0.00	13.33	10.00
NA	20.00	20.00	46.67	46.67	16.67	100.00	30.00	30.00
FALSE	76.67	50.00	50.00	46.67	83.33	0.00	56.67	60.00

Table 24: Overlap of the top 30 PMI emotion holder bigrams with EmoLex in %

Another point worth making is that due to data sparseness for S causes of surprise, disgust, and anger (see table 8), PMI will frequently not be able to determine reliable candidates, particularly regarding direct objects, which are only present in a subset of the instances as depicted in table 18.

4.3.1.1 *Emotion holder*

As can be seen in table 24, emotion holder bigrams frequently are not associated with the emotion in EmoLex.

This does not surprise, as the experiencer does not generally presuppose an emotion, in contrast to the cause of the emotion. The superior precision for anger is due to the fact that *provoke*, the most productive anger pattern, in news frequently refers to inanimate concepts that are – in fact – associated with anger, such as *international outcry*, *angry response*, or *international outrage*. Regarding this evaluation, this is negligible as we focus on causes; in future work, we would want to restrict these patterns by requiring that the emotion holder be animate.

4.3.1.2 *S cause subject + predicate*

The low values for anger and disgust in table 25 are generally due to data sparseness.

EmoLex only inadequately captures the objects of anticipation that are most prominent in news stories: While *grow* as in *gross grow* appears in EmoLex, synonyms like *rise* or *expand* do not. Many fear bigrams can be attributed to viruses and war. While *flu* and *disease* are associated with fear in EmoLex, *virus* is just associated with a negative sentiment. EmoLex captures bigrams such as *flu mutate* or *war destabilize*, but does not account for equally fearful ones such as

Overlap	anger		anticipation		disgust		fear	
	Emo	*Sen*	*Emo*	*Sen*	*Emo*	*Sen*	*Emo*	*Sen*
TRUE	3.33	23.33	33.33	0.00	6.67	13.33	23.33	23.33
NA	16.67	16.67	3.33	100.00	50.00	50.00	36.67	36.67
FALSE	80.00	60.00	63.33	0.00	43.33	36.67	40.00	40.00
Overlap	**joy**		**sadness**		**surprise**		**trust**	
	Emo	*Sen*	*Emo*	*Sen*	*Emo*	*Sen*	*Emo*	*Sen*
TRUE	20.00	36.67	23.33	33.33	3.33	0.00	23.33	36.67
NA	30.00	30.00	6.67	6.67	30.00	100.00	23.33	23.33
FALSE	50.00	33.33	70.00	60.00	66.67	0.00	53.33	40.00

Table 25: Overlap of the top 30 PMI S cause subject + predicate bigrams with EmoLex in %

h5n1 mutate, *virus mutate*, *strain mutate*, or *bird mutate*. Sports-based joy bigrams such as *team achieve* or *player accomplish* are captured by EmoLex, while others that are only joyful because of their compositionality, e.g. *nobody hurt*, *nobody kill(ed)* are not expressed. Furthermore, our approach captures a lot of terms that are very relevant in politics, such as *condition stabilise*. A lot of combinations of subjects and predicates are context-dependent. *Situation occur* or *position arise* might be neutral, positive, or negative, but are associated with sadness in the news domain. Suprise bigrams derived from the S cause subject and predicate are very ambiguous, due to the neutrality of surprise. Bigrams like *settlement allow* or *anyone object* depend on the context more than anything else to elicit surprise. Trust bigrams usually refer to the actions of political leaders or powers such as *democrats handle*, *democrats deal*, or *political contribute*. While a few of these are contained in EmoLex, most of them express a partisan trust.

4.3.1.3 *S cause predicate + object*

A combination of a predicate and its object is more adequate in evoking a sentiment, as evidenced by table 26.

While a few unigrams such as *deny* in *deny honor* are contained in EmoLex, other objects of anger such as *kill teenager*, *refuse plea*, or *molest boy* are not expressed. Some of the highest-ranked bigrams are very context-dependent. For instance, *offer aid* is only a source of anger if is offered virtually unconditionally to a highly controversial entity such as Pyongyang. Even though many bigrams could be correctly considered as source of anticipation, e.g. *rise percent*, *increase percent*, *fall percent*, *reach dollar*, this economics-based anticipation is not captured by EmoLex. Due to data sparseness, we only have a few disgust bigrams, which are very context-dependent.

Overlap	anger		anticipation		disgust		fear	
	Emo	*Sen*	*Emo*	*Sen*	*Emo*	*Sen*	*Emo*	*Sen*
TRUE	16.67	40.00	3.33	0.00	10.00	33.33	33.33	33.33
NA	6.67	6.67	0.00	100.00	10.00	10.00	13.33	13.33
FALSE	76.67	53.33	96.67	0.00	80.00	56.67	53.33	53.33
Overlap	**joy**		**sadness**		**surprise**		**trust**	
	Emo	*Sen*	*Emo*	*Sen*	*Emo*	*Sen*	*Emo*	*Sen*
TRUE	3.33	23.33	30.00	40.00	3.33	0.00	23.33	40.00
NA	3.33	3.33	3.33	3.33	0.00	100.00	3.33	3.33
FALSE	93.33	73.33	66.67	56.67	96.67	0.00	73.33	56.67

Table 26: Overlap of the top 30 PMI S cause predicate + object bigrams with EmoLex in %

Some fear bigrams like *destabilize region*, *upset balance*, or *engulf region* are not expressed in EmoLex, while others like *trigger pandemic* or *spark violence* are, mostly due to their frightening object. Furthermore, others such as *spark (arms) race* or *use (nuclear) program* showcase that more relevant information is needed to satisfactorily determine the relevant emotion without context. Joy bigrams have a very low precision, as unigrams like *survive* in *survive appeal*, *survive attempt*; *escape* in *escape card*, *escape injury*; or *win* in *win set* or *win decision* are not associated with joy in EmoLex, even though they would be a source of joy in the respective bigrams.

Sadness achieves a higher precision, as bigrams like *light candle*, *announce death*, or *leave friend* are labeled correctly. Others, like *wear skirt* are culturally sensitive or context-dependent. Surprise bigrams are highly ambiguous: While *terminate agent*, *include flaw*, *know scientist*, *exclude mortgage*, or *support opposed* might surprise the respective parties, they are not labeled as such in EmoLex. The trust bigrams reduce to trusting someone with a certain action, often to *handle* a *problem*, an *issue*, a *matter*, *security*, etc. These are not captured in EmoLex.

4.3.1.4 *NP cause*

We get the highest agreement numbers overall for bigrams derived from the NP cause, which can be viewed in table 27.

Anger bigrams have the lowest precision; the object of anger often is a body part, e.g. *left knee*, *sore right*, *right ankle* that is bothering professional athletes. We encounter a lot of bigrams of financial anticipation, e.g. *profit of:billion_yen*, *operating loss*, *net loss*, *earnings per:share* that are not captured by EmoLex. This dichotomy of *profit* and *loss* that can be both a source of anticipation showcases the underlying ambiguity of anticipation. A lot of disgust bigrams such as *discovery of:uranium_enrichment_equipment*, *disproportionate use*, *extreme partisan-*

Overlap	**anger**		**anticipation**		**disgust**		**fear**	
	Emo	*Sen*	*Emo*	*Sen*	*Emo*	*Sen*	*Emo*	*Sen*
TRUE	6.67	16.67	13.33	0.00	16.67	43.33	60.00	60.00
NA	33.33	33.33	6.67	100.00	13.33	13.33	16.67	16.67
FALSE	60.00	50.00	80.00	0.00	70.00	43.33	23.33	23.33
Overlap	**joy**		**sadness**		**surprise**		**trust**	
	Emo	*Sen*	*Emo*	*Sen*	*Emo*	*Sen*	*Emo*	*Sen*
TRUE	33.33	73.33	83.33	86.67	16.67	0.00	33.33	46.67
NA	3.33	3.33	0.00	0.00	50.00	100.00	6.67	6.67
FALSE	63.33	23.33	16.67	13.33	33.33	0.00	60.00	46.67

Table 27: Overlap of the top 30 PMI NP cause bigrams with EmoLex in %

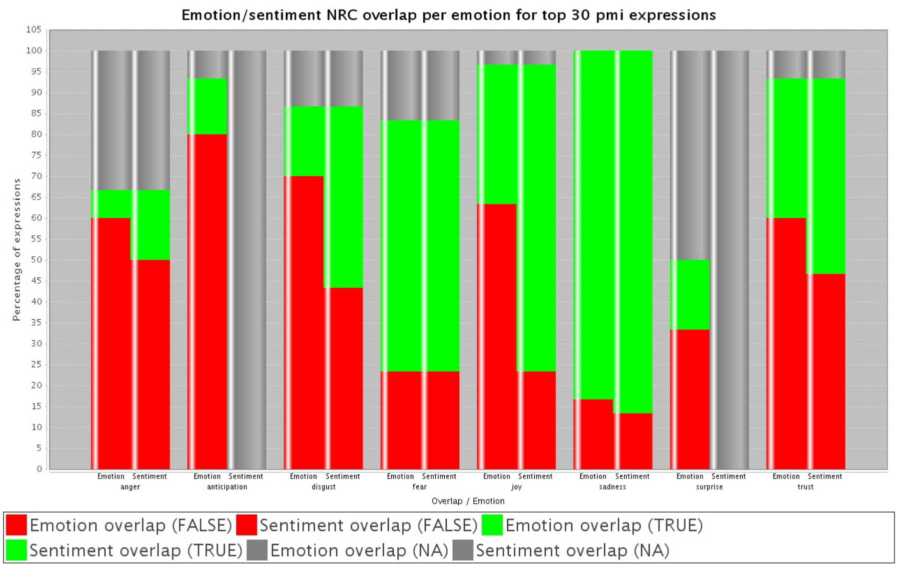

Figure 11: Emotion and sentiment overlap of NP cause bigrams with the NRC Emotion Lexicon

ship, *killing of:civilian*, *path of:violence*, etc. are not captured by EmoLex. Due to news citing involved parties, bigrams that do not mirror the majority's opinion are ranked highly as well, including the dichotomous pair *white people* and *black people*. The same applies to an *international sporting event* that is only included because it was vehemently shunned by North Korea. Significantly, many more of our disgust bigrams are associated with a negative sentiment than are labeled with disgust in EmoLex as can be seen in figure 11, showing that without context, people more frequently have a vague feeling of negativity rather than an intense emotion such as disgust towards an unigram concept.

In contrast, all our fear bigrams that are associated with fear in EmoLex are also labelled with a negative sentiment. Many – like *government reprisal*, *attack by:ethnic_albanian*, *government retribution*, *revenge attack* – are identified correctly using EmoLex, while others like *political backlash*, *long-term effect*, *unintended consequence*, *new crackdown*, etc. are not identified. Our joy bigrams deal with joy particularly in international relations and politics. While bigrams like *diplomatic immunity*, *widespread support*, *comfortable lead*, *high popularity* are not labeled with joy in EmoLex, they are labeled as positive. However, bigrams like *equal rights*, *wide support*, or *best season* are not even labeled as positive. For sadness, we generally receive the highest agreement numbers as well as the highest PMI scores. They can be seen in table 19. Without excluding named entities, entities like *Jean-Marie Le Pen*, a far-right French politician or Walt Disney – which has a track record of surprising audiences as well as investors – are among the top-ranked bigrams for surprise. As surprise in news usually is large-scale, disastrous events like an *apparent suicide*, a *powerful earthquake*, and a *brutal* or *double murder* are also ranked highly. Finally, among the trust bigrams derived from the NP cause, we find many concepts that people or countries rely on, such as *private insurance*, *computer model(s)*, *public transportation*, *fossil fuel*, *immigrant labor*, *subsistence farming*, *foreign worker(s)*, etc. We also get glimpses of objects of trust relevant for particular groups, e.g. sailors trust their *(global) positioning system*, while tennis players trust a *strong serve*.

4.3.1.5 *Summary*

As we have mentioned initially, EmoLex is only partially reliable to validate our results due to the compositionality of emotion. However, it serves to establish that bigrams derived from the NP cause displayed in figure 11 are generally most evidently associated with a certain emotion, while a combination of a predicate and its object ranks second. Furthermore, it shows that for anger, disgust, and surprise, data sparseness for clausal causes is indeed a problem. However, even if given enough data, disgust is inherently more difficult to assign than fear. An initial investigation of our results has shown that the positive joy and trust and the negative sadness and fear contain the most reliable concepts. Many of them also contain valuable information about certain groups and should be considered with the relevant context of emotion holder and prepositional objects.

4.3.2 *Evaluation using manual annotations*

EmoLex has provided us with a first guideline for our evaluation. In order to further validate our results, we create a gold standard by letting annotators manually label bigrams of the NP cause and the predicate + object of the S cause as these have produced the best results in the previous evaluation.

Mohammad and Turney include a question like the following in their annotation task in order to 'anchor' the intended word sense and retrieve the prior probability of that word sense's emotion association.

Q. Which word is closest in meaning (most related) to *startle*?

- *auto mobile*
- *shake*
- *honesty*
- *entertain*

The bigram characteristic of our terms already anchors the word sense in most cases. However, the emotion often is still context-dependent. For this reason, we include a randomly selected sentence from which the bigram was extracted with the respective emotion in order to verify that the bigram would carry the same emotion with and without context.

As we have seen during the EmoLex evaluation, not all bigrams are associated with an emotion without context. Thus, in order not to bias the annotator towards an association with a particular emotion, we initially pose the question whether the concept is associated with an emotion at all and if so, we ask for specification.

Below you can find the annotation questions for the target concept *announce death* and a randomly selected sentence:

Concept: *announce death*

Q1. Is the concept associated with an emotion?

- yes
- no

Q2. If so, with which emotion is it associated?

- anger
- anticipation
- disgust
- fear
- joy
- sadness
- surprise
- trust

Sentence: *" the federal government regrets to announce the sudden death of chief m.k.o. moshood abiola , " the statement said .*

Q3. Is the concept associated with an emotion in the sentence?

- yes
- no

Annotated bigrams	**Number**
Unanimous emotions	157
Majority emotions	271
Unanimous sentiment	185
Majority sentiment	309
Total	320
Fleiss' κ	0.42

Table 28: Number of annotated bigrams for emotion and sentiment agreement

Q4. If so, with which emotion is it associated?

- anger
- anticipation
- disgust
- fear
- joy
- sadness
- surprise
- trust

We exclude all named entitites from the bigrams, since annotators might not be familiar with them and their mention introduces an increased dependence on context. In order to minimize the expenditure of time spent annotating while still guaranteeing reliable results, we select the top 20 bigrams per emotion for the NP cause and the concatenated predicate and object of the S cause. This leaves us with 320 bigrams, which we hand off to three annotators (including the author).

We provide annotation guidelines, examples highlighting possible annotations, and Plutchik's emotion wheel as a reference for annotation.

Our annotators achieve a Fleiss' κ of 0.42, which signifies moderate agreement [20] (cf. table 4). This is low in comparison to other tasks such as part-of-speech tagging but significantly higher than the agreement of 0.29 observed by Mohammad and Turney who performed a similar annotation task. This can be attributed to the selection process of our bigrams, which increased their prior probability of being associated with an emotion, thereby facilitating agreement. Further statistics can be taken from table 28: 49% of all bigrams have been unanimously annotated with an emotion, while a majority agreed on the emotion of 85% of bigrams. For sentiment, we observe unanimous agreement of 58% and majority agreement of 97% respectively.

We will use these 271 bigrams and 309 bigrams on which the majority agreed in the following as gold standard for emotion and senti-

Emotion/sentiment	**Number**
Anger	0
Anticipation	0
Disgust	2
Fear	16
Joy	44
Sadness	41
Surprise	0
Trust	1
None	167
Total	271
Positive	51
Negative	91
Neutral	0
None	167
Total	309

Table 29: Majority emotion/sentiment distribution of bigrams

ment respectively. Table 29 depicts the distribution of these bigrams. Quite significantly, even though the guidelines encouraged to consider all emotions carefully, anger, anticipation, and surprise have never been selected by the majority; disgust and trust have only been chosen twice and once respectively. In our own annotations, we have observed that these kinds of emotions are more strongly associated with personal opinions. These opinions, we found, often require more context to fully manifest themselves, thus letting an undefined, underlying mood of fear, joy, or sadness prevail. Indeed, many disgust bigrams have been labelled as indicative of sadness by annotators. This is also a result of requesting the annotators to assign only the most prominent emotion, as emotions such as anger or disgust might be present to a degree and could have been included as a second choice.

These results are in line with our previous investigations showing that bigrams without context frequently are not very indicative of anger, anticipation, and surprise. Furthermore, if a bigram was connoted negatively, annotators would often associate different negative emotions with it, rendering the negative majority agreement significantly higher than the majority agreement of negative emotions combined. These difficulties account for the comparatively low number of bigrams that have been labeled by the majority of annotators with an emotion (38%) and a sentiment (45%) respectively. These percentages, though, agree with the 22.5% of emotive terms reported by Mohammad and Turney.

	NP cause			**S cause predicate + object**		
Emotion/sentiment	*Precision*	*Recall*	*F1*	*Precision*	*Recall*	*F1*
anger	0.00	-	-	0.00	-	-
anticipation	0.00	-	-	0.00	-	-
disgust	0.07	1.00	0.13	0.06	1.00	0.11
fear	0.33	1.00	0.50	0.40	0.89	0.55
joy	0.90	0.95	0.92	0.83	0.60	0.70
sadness	1.00	0.67	0.80	0.36	0.45	0.40
surprise	0.00	-	-	0.00	-	-
trust	0.00	-	-	0.05	1.00	0.10
total – emotion	0.36	1.00	0.53	0.23	1.00	0.37
positive	0.56	0.96	0.71	0.40	0.57	0.47
negative	0.60	0.84	0.70	0.37	0.87	0.51
neutral	0.00	-	-	0.00	-	-
total – sentiment	0.47	1.00	0.64	0.29	1.00	0.45

Table 30: Precision, recall, and *F1* score for emotions/sentiments of NP cause and S cause predicate + object bigrams

We have listed precision, recall, and *F1* score per emotion and sentiment and by NP cause and S cause predicate + object of our predicted annotations against the gold standard of manually annotated bigrams in table 30.

Due to the non-existent annotations for anger, anticipation, and surprise, precision for these emotions is at 0.00, and for trust at 0.05. All annotations labelled by the annotators as disgust were also predicted as disgust, though we obtain many false positives. The S cause bigrams have a higher recall and precision for fear than the NP cause bigrams. Joy achieves good precision and recall for both NP and S cause. While all NP cause bigrams that have been predicted as sadness also have been labelled by the annotators with sadness, many more have been labelled by the annotators that PMI assigned to an emotion other than sadness. In contrast, the performance of S cause bigrams for sadness is lower, showing that a concept or event might more unambigously indicate sadness than an action.

In total, we obtain a better performance for bigrams derived from the NP cause than from the predicate and object of the S cause, though the difference in the *F1* score is not as significant as the initial evaluation with EmoLex might have suggested.

Concerning sentiment, precision is 0.00 for neutral sentiment. As expected, we achieve a higher *F1* score for sentiment than for emotion, with NP cause bigrams being more accurate than predicate + object of the S cause. For the NP cause, all 51 positively annotated bigrams have been predicted correctly, while the many false positives decrease the precision. As annotators have identified bigrams as negative more often, the predicted negative sentiment corresponds more frequently to the annotated sentiment leading to a higher precision.

In the following, we will look at the individual emotions in more detail, to derive further lessons and insights from the evaluation.

4.3.2.1 *Anger*

As we have already mentioned, a lot of anger terms are context-dependent. All of our NP cause anger bigrams are assigned no emotion by annotators, while some S cause bigrams are assigned different emotions, e.g. *accept emmy* is labelled with joy, while the term was scored highly by PMI because it was associated on two different occasions with anger in the context of a certain *off-color (emmy acceptance) speech* (`APW_ENG_20070920.1099/0`, `APW_ENG_20070919.1625/0`), while not appearing with any other emotions. This shows that, while our discounting method is able to filter out ngrams that appear only once, it fails given the lack of more promising candidates. The underlying source of this issue is the comparatively low amount of data: Only 7,207 S causes are associated with anger (cf. table 8), while only a subset of these contain direct objects (cf. table 18).

4.3.2.2 *Anticipation*

Anticipation NP bigrams have been mostly annotated with no emotion as a lot of them deal with money, which is inherently neutral, e.g. *earnings per:share*, *revenue of:$*, *cent per:share*, etc. A high-scoring bigram, *map in:color*, is due to a parsing error, as *forecast* in *National forecast map in color* has been continuously misparsed as a verb. Some anticipation bigrams referring to *revenue* or *operating loss* have been labelled with sadness. Anticipation predicates and direct objects have been predominantly annotated with no emotion, with only *see growth* and *post profit* having been labelled with joy.

4.3.2.3 *Disgust*

Among disgust NP cause bigrams, *brutal persecution* has been labelled with disgust by the annotators. Others have either been labelled with no emotion – as the expression of disgust mostly depends on context – or with sadness, e.g. *continued failure*, *path of:violence*, *killing of:civilian*, showcasing the overlap between disgust and sadness and the bias towards sadness. For disgust S cause bigrams, *clergy abuse* has been labelled with disgust by the annotators. Again, others have been labelled either with no emotion or joy. While *buy ring* is generally perceived as positive, *Jerry Seinfeld only (buying) a two-carat ring for his new bride* (`NYT_ENG_20000212.0066/27`) evokes disgust in insiders. These outliers achieve dominance due to lack of data failing to produce better candidates given the only 544 S causes associated with disgust (cf. table 8).

4.3.2.4 *Fear*

For fear NP causes, we achieve a moderate precision, with *possible reprisal*, *attack by:ethnic_albanian*, or *humanitarian disaster* having been labelled with fear by the annotators. For others, such as *security vacuum*, *new crackdown*, *long-term effect*, association with fear is of a more covert and underlying nature, as they have been labelled with no emotion by the annotators. Finally, *loss of:business* or *loss of:sovereignty* have been been labelled with sadness, showcasing a clear bias towards sadness. We achieve a perfect recall, as no other bigrams have been labelled with fear by the annotators. Significantly, fear is the only emotion for which S cause bigrams achieve a higher performance than NP cause bigrams. *Trigger pandemic*, *destabilize region*, *spark pandemic*, *upset balance*, *spark violence*, *engulf region*, and *trigger violence* have all been labelled by the annotators as clear indicators of fear. Others were either labelled with no emotion, e.g. *use (uranium) enrichment*, *set precedent*, *spark (weapon) race*, or with sadness, e.g. *disrupt export*, *hurt industry*.

4.3.2.5 *Joy*

Joy NP cause bigrams have almost all been labelled with joy by the annotators, showcasing that concepts like *widespread support*, *traditional friendship*, *excellent relation*, *equal rights*, *high popularity*, *comfortable lead*, etc. are clear indicators of joy. Only *diplomatic immunity* and *degree of:autonomy* have been labelled with no emotion, as their association with joy is more subtle. Similarly, we receive a high precision for S cause bigrams, as bigrams such as *win set*, *survive attempt*, *escape injury*, *watch movie*, *win decision*, *avoid injury*, *have insurance* are seen by annotators as being clearly associated with joy. Only *read history*, *get draw*, and *have guy* have been labelled with no emotion, as joy for these is more situation-specific: some people enjoy reading about history, while others are lucky to get a good draw or to have a good guy on the team.

4.3.2.6 *Sadness*

All sadness NP cause bigrams – the top 10 can be seen in table 19 – have been labelled with sadness by the annotators, confirming that sadness is both a strong and unambiguous emotion. The lower recall is due to the observed bias towards sadness, with bigrams of other emotions having been labelled with sadness by the annotators. Conversely, S cause bigrams achieve a significantly lower precision. Bigrams such as *announce death*, *create anguish*, *leave friend*, *overshadow day*, and *make weep* have been labelled by annotators with sadness, while the sadness connotation of others was too covert or context-dependent, leading them to label bigrams such as *light candle*, *steal cut*, *prevent compatriot*, etc. with no emotion.

4.3.2.7 *Surprise*

For surprise NP cause bigrams, a majority agreed on only 7 of 20 bigrams, showcasing the ambiguousness of surprise. All – except for *apparent suicide*, which was labelled with sadness – were assigned no emotion, confirming the dependence on context of surprise bigrams that we previously observed. Similarly, surprise S cause bigrams have been labelled predominantly with no emotion, with only *terminate agent* and *include flaw*, and *reverse impairment* having been assigned other emotions, sadness and joy, respectively.

4.3.2.8 *Trust*

Significantly, all trust NP and S cause bigrams have been assigned no emotion. The association of nominal phrases such as *public transportation*, *private insurance*, and *computer model*, and of predicates and objects such as *handle issue*, *meet payment*, and *implement deal* with trust seems to be too weak to warrant a clear label. Trust as the only other clearly positive emotion besides joy seems to be significantly weaker than its counterpart, justifying its exclusion from other emotion typologies such as Ekman's.

4.3.2.9 *Sentiment*

The performance regarding sentiment reflects the performance of the individual emotions, with trust and anger and disgust decreasing the precision of positive and negative sentiment respectively. Still, we achieve a good precision for positive and negative sentiment of 0.56 and 0.60 respectively, that – combined with an excellent recall – amounts to a solid *F1* score of 0.71 and 0.70 respectively for positive and negative sentiment for nominal causes. For clausal causes, we achieve lower *F1* scores of 0.47 and 0.51 for positive and negative sentiment respectively. As no bigrams have been labelled with a neutral sentiment, our precision for neutral sentiment is 0.

4.3.2.10 *Sentences*

On a sentential level, annotators were predominantly able to identify the correct emotion. Some misparses made annotation more difficult, e.g. *hate* in *A bid to pass hate crime legislation after the 1998 attack died in the Texas legislature without Bush's support* was misparsed as a verb, causing *crime legislation* to be labelled as a cause.

One of our annotators often did not identify *count on* and *rely on* as indicators of trust, and sometimes did not consider *harass* an indicator of anger, or *shock* an indicator of surprise. The same annotator did not associate *expect* and *predict* in the context of analyst forecasts with anticipation. These idiosyncrasies showcase that association with emotion – even as we try to gauge it empirically – remains a subjective impression that is inherently vague and differs from person to person. For our next annotation task, we will employ more annotators to compensate these small differences.

Annotators observed an unemotive connotation of *enjoy* occurring in sentences like *China has adopted various measures to ensure that the disabled enjoy equal rights with other citizens* (`XIN_ENG_19951227.0085/4`), while one annotator argued that *be lucky to* in *Their driver and interpreter where lucky enough to escape with minor injuries* (`XIN_ENG_20050603.0017/24`) is not associated with joy due to the informative style of the writing. These issues related to the ambiguity of emotions pose new challenges that need to be considered when expanding the patterns or applying them to a new domain.

4.3.2.11 *Summary*

All things considered, our bigrams weighted with PMI produce reliable results for fear and excellent results for joy and sadness. For these emotions, the lists that we have provided[7] can be used as emotion lexicons. As we have remarked, they provide added context over existing resources that contain unigrams and are particularly valuable in the news domains. For the other emotions, due to data sparseness, emotion selection bias, and dependence on context information, the prior probability of association is not high enough to guarantee dependable outcomes. While they can be used in applications, additional context features such as context-windows are advisable to inform analyses.

4.4 TOPIC MODELLING

So far, we have considered emotion-evoking expression mostly in isolation. We have identified commonalities between expressions, e.g. that many ambiguous concepts in table 21 refer to state leaders and states or that loss and death are common themes among sadness bigrams in table 19. However, we derived these insights only through manual review. Furthermore, a recurring theme of this project has been that emotions depend – at least partially – on their context. By observing solely unigrams and bigrams, we have omitted part of this context that might be relevant for the disambiguation of emotion association. Latent Dirichlet Allocation (LDA) [6] also known as *topic models* has become the de facto standard for identifying semantic structure in documents in the field of document modelling. We will use

7 The PMI lists can be found at https://github.com/sebastianruder/emotion_proposition_store/tree/master/out/scores/pmi.

it in this section to investigate the two aforementioned issues: (a) We will explore underlying topics that are associated with an emotion; and (b) we will surface additional emotion-evoking expressions by taking into account additional context.

4.4.1 *LDA*

LDA is a generative probabilistic model of a corpus that aids in discovering latent semantic themes in text. Similar models suffer from the following two weaknesses: (a) The *mixture of unigrams* model [33], which represents documents by drawing words of every document from a conditional multinomial distribution with a discrete random variable, limits each document to one topic; and (b) they do not generalize easily to new documents in the case of probabilistic latent semantic indexing (pLSI) [18].

LDA mitigates these deficits by allowing documents to exhibit multiple topics to different degrees and by being able to assign topic probabilities to unseen documents. It represents documents as a mixture of an underlying set of latent topics, where each topic is characterized by a distribution of words.

Blei et al. show the usability of their work for document modelling, collaborative filtering, and document classification. In the latter, they use an SVM trained on features induced by a 50-topic LDA model achieving a dimensionality reduction of the feature space by 99.6% in comparison to a simple bag-of-words model and greater accuracy.

4.4.2 *Using LDA for modeling lexical semantics*

LDA was originally conceived and is typically used in a fully unsupervised way, with only the number of topics being specified in advance. LDA represents these topics as latent variables and estimates the topic distributions over documents θ_d and topic-word distributions ϕ_t. By estimating these parameters on a document collection, we obtain topic proportions $P(t|d)$, i.e. the probability of a topic given a document, and topic distributions $P(w|t)$, i.e. the probability of a word given a topic. These can be used to compute a smooth probability distribution $P(w|d)$ of words given a document as in equation 10, where t denotes a latent topic, w a word, and d a document in the corpus.

$$P(w|d) = \sum_t P(w|t)P(t|d) \tag{10}$$

Recently, LDA has not only been employed for unsupervised document modelling, but also for inducing semantic knowledge from high-dimensional co-occurrence data in a semi-supervised way. Ritter et al. and Séaghdha use LDA to model selectional verb preferences, while Hartung and Frank apply LDA to attribute selection. Both collect pseudo-documents, i.e. bags of words containing co-occurrence data, to induce topic distributions characterizing observed topic mixtures. For verb preferences, these pseudo-documents consist of co-occurring syntactic arguments, while for attribute selection, adjectives and nouns that co-occur with the attribute nouns in local contextual relations are used.

We apply Controlled LDA (C-LDA), an extension employed by Hartung and Frank that allows to take supervised category information into account, to emotion detection. For this purpose, we construct pseudo-documents that characterize Plutchik's eight emotions by collecting causes that co-occur, i.e. are associated with, said emotion. By fitting an LDA model to the collection of these pseudo-documents, we are able to obtain topic distributions that are most closely associated with specific emotions. An investigation of these reveals new insights into underlying themes pertaining to individual emotions or shared across several emotions.

4.4.3 *Controlled LDA*

C-LDA functions similarly to regular LDA. They solely differ in the nature of documents used for training the topic models: While LDA is usually trained on a natural language corpus, C-LDA presupposes a supervised structuring of the input documents so that they convey semantic information that characterizes the required categories, in our case emotions.

Populating the pseudo-documents with the highest-ranking NP cause and S cause bigrams would allow us to detect underlying themes in the emotion-evoking expressions; we would, however, fail to capture context that might provide additional insights. The bags of word of the leaves of our causes include subordinate clauses (cf. example 5, section 3.8) that provide additional context and are thus ideal as the basis of our topic models. We select the bag-of-words representation of the top 200 highest-scoring NP and S subject + predicate cause bigrams for each emotion and aggregate them in a pseudo-document for their respective emotion. This way, we obtain eight pseudo-documents corresponding to Plutchik's eight emotions. Their token and type counts can be found in table 31.

We obtain high token and type counts for anticipation, joy, and fear, while surprise, anger, and disgust yield predictably low counts. These results are in line with the distribution of NP and S causes that we observed in table 8, as clausal causes typically contain significantly more tokens than nominal phrases.

By presenting LDA with these pseudo-documents, we are able to add supervision through linking our emotion categories to the gen-

Emotion	# of tokens	# of types
disgust	11,967	1,451
joy	110,679	7,368
sadness	21,134	2,504
surprise	5,922	1,308
anticipation	1,025,895	22,342
trust	22,205	2,963
fear	92,399	6,408
anger	9,745	1,762

Table 31: Token and type ratios for the emotion pseudo-documents

erative process. As we assign each pseudo-document to an emotion, the content of such a pseudo-document can be thought to represent the association with said emotion. In line with the distributional hypothesis, which states that "a word is characterized by the company it keeps" [15], these pseudo-documents can be seen as distributional fingerprints of the association with an emotion. Thus, we approximate $P(w|e)$, the probability of a word given an emotion, by $P(w|d)$, the probability of a word given a document, as obtained from LDA:

$$P(w|e) \approx P(w|d) = \sum_t P(w|t)P(t|d) \tag{11}$$

We rely on MALLET [24] for implementation of C-LDA. We run 1,000 iterations of Gibbs sampling using default values for all hyperparameters.

Topic models allow us to view the latent topics of a text in different granularities. While our pseudo-documents have been structured according to eight emotion categories, there exist clearly more than these 8 topics, e.g. narrower topics for one emotion or topics that overlap between different emotions. For this purpose, we experiment with different configurations, generating topic models for 10, 20, 30, and 50 topics respectively. For each emotion pseudo-document, we take the topic with the highest topic proportion $P(t|d)$ to be associated with the corresponding emotion. For each of these eight topics, we output the 20 words with the highest probability given the respective topic, i.e. $P(w|t)$. These words maximize equation 11 and thus must also have the highest probability given an emotion, i.e. $P(w|e)$.

For an initial comparison, we show the top 7 key words for the 10-, 20-, and 30-topic configurations in table 32.[8]

Notably, the key words for the topic most associated with a particular emotion do not differ much among configurations of 10, 20,

8 We do not show the key words for the 50-topic configuration as they are mostly identical.

Emotion	**# of topics**	**Key words**
disgust	10	aid food force government international foreign violence
	20	force violence government people act recent civilian
	30	force violence government act people recent international
joy	10	support relation good widespread rating popularity lead
	20	support relation good widespread rating popularity wide
	30	support relation good widespread rating popularity wide
sadness	10	loss life death civilian innocent tragic victim
	20	loss life death innocent civilian tragic victim
	30	loss life death innocent civilian tragic victim
surprise	10	state world nation leader year country make
	20	shock news stun murder world communist leader
	30	shock stun news murder world state communist
anticipation	10	percent billion profit yen net share cent
	20	percent billion year profit yen million share
	30	percent year profit yen net million growth
trust	10	aid food force government international foreign violence
	20	aid food power foreign private international nuclear
	30	aid food power foreign private international import
fear	10	violence war nuclear give spark region attack
	20	violence nuclear spark give trigger lose undermine
	30	spark attack region trigger lose undermine political
anger	10	iranian beijing left president chinese chen boat
	20	iranian left chen boat knee beijing pro-independence
	30	iranian left chen knee pro-independence boat beijing

Table 32: Top 7 key words for the topic most associated with an emotion for different number of topics

Overlap	anger		anticipation		disgust		fear	
	Emo	*Sen*	*Emo*	*Sen*	*Emo*	*Sen*	*Emo*	*Sen*
TRUE	8.00	12.00	10.00	0.00	8.00	28.00	28.00	32.00
NA	34.00	34.00	18.00	100.00	22.00	22.00	32.00	32.00
FALSE	58.00	54.00	72.00	0.00	70.00	50.00	40.00	36.00
Overlap	**joy**		**sadness**		**surprise**		**trust**	
	Emo	*Sen*	*Emo*	*Sen*	*Emo*	*Sen*	*Emo*	*Sen*
TRUE	18.00	40.00	26.00	30.00	10.00	0.00	16.00	22.00
NA	16.00	16.00	12.00	12.00	32.00	100.00	10.00	10.00
FALSE	66.00	44.00	62.00	58.00	58.00	0.00	74.00	68.00

Table 33: Agreement numbers with EmoLex for the top 50 key words for the topic most associated with an emotion

30, and 50 emotions. We focus on the 20-topic LDA configuration for initial further observations as it strikes a good balance between a lack and an abundance of topics. To gain a rough overview of the kinds of words that define these topics, we turn to EmoLex for another evaluation. We compare the emotion association of the top 50 key words for the most probable topic for an emotion with EmoLex using the methodology we described in section and show the results in table 33.

Agreement numbers are slightly lower than for NP causes depicted in table 27, but show that topics for joy, sadness, and fear are the most salient topics. Emotion agreement for joy is lower than for sadness and fear, but sentiment agreement is higher, showing that many joy key words are associated with a positive sentiment rather than explicitly indicate joy.

As we can see in table 34, the most probable topic for an emotion has a significantly higher probability than the second most probable topic. Only joy, sadness, and fear contain another topic that has a topic proportion > 0.1. These second-ranked topics have the following key words:

joy: party political support people win health include government house public democratic family man

sadness: state country make election world president east month peace south power year

fear: war attack security global military people china law rebel move process islamic

In the following, we summarize the essence and the points of overlap of the most associated topics for joy, sadness, and fear (see table 32) as well as of their second-ranked topics with a topic proportion > 0.1 depicted above. While the most associated joy topic can be thought of as *political campaigns*, the second one can be conceived as

Emotion	Topic #1 probability	Topic #2 probability	Topic #3 probability
disgust	0.82	0.07	0.03
joy	0.65	0.12	0.06
sadness	0.67	0.11	0.06
surprise	0.73	0.07	0.04
anticipation	0.56	0.08	0.07
trust	0.67	0.07	0.05
fear	0.48	0.20	0.10
anger	0.67	0.08	0.06

Table 34: The topic proportions of the top 2 most probable topics of 20 topics given a pseudo-document

government and values and also ranks second for surprise and trust. The first sadness topic can be labelled as *tragedies*, while the second one can be summarized as *state actions*, which ranks second also for surprise and anger, and third for disgust and fear, showing a clear overlap between these emotions. Finally, the first-ranked fear topic can be described as *international violence and its consequences*, while the second one encompasses *war and military*, ranking second for disgust and third for anger, surfacing the similarities between these emotions.

These 20 topics, while exposing the most prominent topic for each emotion, only adumbrate the diversity of underlying topics hidden in the documents. To gain a better understanding of their diverse nature, we take a look at the top three topics for each emotion from the 50-topic configuration. In the following, we state the top 3 topics for each emotion together with their most salient key words.

A. disgust
   a) *violence against civilians* (force, violence, government),
   b) *attacks undermining security* (attack, police, security),
   c) *military government actions or elections* (government);

B. joy
   a) *political campaigns* (support, relation, good, widespread),
   b) *economic status and personal prestige* (economic, status),
   c) *rights and family* (freedom, autonomy, rights, home, family);

C. sadness
   a) *tragedies* (life, loss, death, innocent),
   b) *conflicts* (leader, israeli, attack, gaza, africa),
   c) *military government actions or elections* (see disgust);

D. surprise

a) *unforeseen events* (news, world, murder, leader, communist),

b) *tragic international events* (death, nation, child, suicide),

c) *military government actions or elections* (see disgust);

E. anticipation

a) *earnings, profit or loss* (all three topics);

F. trust

a) *international dependencies* (food, private, power, import),

b) *political groups* (support, day, union, member, opposition),

c) *monetary reliance* (income, employer, payment, borrowing);

G. fear

a) *international violence and its consequences* (violence, spark),

b) *military government actions or elections* (see disgust),

c) *outbreaks of conflict or disease* (pandemic, conflict, outbreak);

H. anger

a) *military harassment and state visits* (chinese, visit, vessel),

b) *political parties* (pro-independence, rhetoric, office, party),

c) *military government actions or elections* (see disgust).

These topics highlight themes most of which we have already observed among the concepts we investigated thus far. Interesting points of focus that we have not encountered are the following: (a) joy that is not related to politics, but to one's own *economic status and personal prestige* as well as fundamental values of society; (b) trust, whose top three topics are clearly distinguished across different domains, ranging from the international scale, over political and communal groups, to personal monetary matters; (c) the topic *military government actions or elections* that is shared among disgust, sadness, surprise, fear, and anger.

In summary, these topics provide a higher-level vantage point than individual ngrams on common themes that are associated with certain emotions.

5 OUTLOOK AND CONCLUSION

We dedicate this final chapter to the dual purpose of prospect and retrospection. In section 5.1, we outline several promising future avenues for investigation and research. Finally, in section 5.2, we reiterate our contributions, summarize our main findings and central themes, and provide a conclusion.

5.1 OUTLOOK

An addition of data has been generally found to improve the reliability both of machine learning algorithms as well as data-dependent measures such as PMI, in our case particularly the PMI scores for clausal cause bigrams for anger, disgust, and surprise. There are two approaches to gathering more data: (a) On the one hand, staying with the existing domain and corpus and learning new patterns using bootstrapping (cf. [35]). During this process, we will additionally refine existing patterns, e.g. by adding an animateness feature as mentioned in section 4.3.1.1, and manually add patterns that refer to clausal causes. This will offer the benefit of expanding our available data, while keeping it relevant to the news domain. (b) Alternatively, we use our patterns to mine data from the web (cf. approaches described in section 2.3). This will make the data more general-domain, thereby increasing its relevancy to other domains, while simultaneously reducing its significance for the news domain.

More data will also improve the language modelling capabilities by allowing us to use more complex ngrams, such as trigrams, which require more data than bigrams to produce relevant results, but have the advantage of taking into account more context.

An LDA involving Dirichlet-multinomials over trigrams rather than unigrams could also be considered in order to render the topic models more context-sensitive and consequently more adequate for language and emotion modelling. Training word vectors for emotion detection rather than sentiment analysis (cf. [23]) would be another interesting research avenue.

Our data also allows an investigation of the temporal aspect of emotions, which would reveal which concepts and actions evoke different emotions over time and how the emotions associated with these concepts changes.

Finally, the compositionality of emotion warrants further attention. What makes certain compounds emotive, while others remain neutral? How can we predict the emotion of an unseen event? As a starting point, an SVM can be trained to predict the emotion label of our bigrams for joy, sadness, and fear – as these are the most reliable – using the emotions of the constituent unigrams in EmoLex [30] as well as context features of the whole sentence in which the bigrams occurred (minus the patterns) as features. Unseen, labelled examples for evaluation can then be mined from the web using our patterns.

5.2 CONCLUSION

In the following, we summarize our four main contributions to the field of emotion detection along with our findings:

A. We designed and evaluated patterns that are frequent and clearly associated with an emotion. These can be used as-is to extract emotion holders and causes from a corpus or the web. A retrospective analysis revealed that certain patterns can appear in domain-dependent unemotive contexts, leading to erroneous instances. A careful selection in the case of manual curation and an accurate reliability measure for bootstrapping thus is key to pattern creation.

B. We have acquired more than 1,700,000 propositions from the Gigaword news corpus using these patterns, filtered, and generalized them. The same pattern-based extraction approach can be applied to generate a domain-specific compositional emotion lexicon from other domains. The capabilities to access additional context as well as to represent events were key considerations regarding this pattern-based extraction process.

C. We have stored these propositions in an Emotion Proposition Store, which we made available to the research community. We showcase the potential of such a resource and outline its use in future applications.

D. We analysed different aspects of the resource: (a) Through a comparison of the point-wise mutual information and chi-square association metrics, two measures commonly used in sentiment analysis, we found that, although both of them are equally apt in determining promising candidates for emotion detection and though chi-square is a normalized measure, PMI is better suited for our purposes as it facilitates comparison across emotions; (b) a comparison of the salience of unigrams and bigrams to evoke emotions revealed that, while unigrams can be connoted with emotions, association is much clearer and less ambiguous for bigrams; (c) an analysis of the most ambiguous and controversial concepts yields state leaders and states as the most controversial entities, which neither show association with anticipation nor fear; we find that anticipation only produces weak association even with its highest-ranked examples, while fear tends not to occur strongly with non-negative emotions.

Furthermore, we generated lists of the highest-ranking unigrams and bigrams that are most associated with an emotion, which can be used as an emotion lexicon for the news domain. The resulting lists are more finely granular than existing resources, distinguishing between emotion holder and different cause types; their bigram representation incorporates more context than existing unigram representations, while additional context can be retrieved via a bag-of-words representation of the cause or cross-reference with the extracted sentences. We reviewed these lists using both an evaluation against an existing emotion lexicon as well as an evaluation against human-annotated data. We find that a nominal cause as well as a combination of predicate and object are most apt to evoke an emotion; regarding emotions, we find anger, anticipation, surprise, and – to a lesser degree – disgust and trust to be difficult to ascertain without context and to be dominated by sadness, fear, and joy, for which we achieve excellent results. Finally, we uncovered underlying themes that are associated with specific emotions using an extension of Latent Dirichlet Allocation; by populating pseudo-documents with contextualized causes, we were able to introduce supervised category information into the generative process to derive prominent topics and topics that are shared between emotions.

The recent advances in sentiment analysis and its growing use in applications pave the way for a more finely granular detection that more closely reflects actual human feelings and emotions. In the future, AI agents will need to be able to detect human emotions in order to bring down conversation barriers, prevent misunderstandings, and facilitate interaction. Emotion detection as a research domain, however, is still in its infancy. Our insights shed light on problems and chances in this field, while our proposed pattern-based extraction method can be replicated to derive knowledge from other domains. Finally, by sharing our data and generated resources, we hope to lay the foundations for further progress in the field and to contribute to advancing the state-of-the-art.

APPENDIX A

A.1 MAJORITY PATTERNS

Emotion	Pattern
anger	harass
anger	offend
anger	irk
anger	frustrate
anger_1	be irate that
anger_1	outrage
anger_1	inflame
anger_1	infuriate
anger_1	be incense that
anger_1	enrage
anger_2	scorn
anger_2	torment
anger_2	be angry at
anger_2	vex
anger_2	be angry that
anger_2	be angry with
anger_2	afflict
anger_2	insult
anger_2	aggravate
anger_2	be angry about
anger_2	fret about
anger_2	rankle
anger_2	provoke
anger_3	bother
anger_3	irritate
anger_3	grate
anger_3	annoy
anger_3	sting
anger_3	disturb
anticipation	yearn for

anticipation	predict that
anticipation	ache for
anticipation	be intent on
anticipation	intrigue
anticipation_2	aspire
anticipation_2	be hopeful that
anticipation_2	expect
anticipation_2	anticipate
anticipation_2	covet
anticipation_2	aspire to
anticipation_2	crave
anticipation_2	foresee
anticipation_2	hunger for
anticipation_2	predict
anticipation_2	prepare for
anticipation_2	long for
anticipation_2	await
anticipation_2	forecast that
anticipation_2	hope for
anticipation_2	expect that
anticipation_2	be eager for
anticipation_2	look forward to
anticipation_2	forecast
anticipation_3	be curious about
disgust	horrify
disgust	shun
disgust_1	abhor
disgust_1	loathe
disgust_1	despise
disgust_1	hate
disgust_1	deplore
disgust_1	disdain
disgust_2	dislike
disgust_2	shame
disgust_2	be ashamed of
disgust_2	alienate
disgust_2	repel
disgust_2	disgust

fear	be afraid of
fear	unnerve
fear	rattle
fear	dread
fear_1	terrify
fear_1	be terrify that
fear_2	be frightened that
fear_2	be scared that
fear_2	fear that
fear_2	frighten
fear_2	fear
fear_2	intimidate
fear_2	spook
fear_2	be scared of
fear_2	scare
fear_2	be afraid that
fear_3	concern
fear_3	be anxious that
fear_3	be anxious about
fear_3	be apprehensive about
fear_3	worry
fear_3	be nervous about
joy	be enthusiastic about
joy	comfort
joy	fulfill
joy	desire
joy_1	thrill
joy_1	be elate that
joy_1	rejoice
joy_1	be ecstatic that
joy_1	be thrill that
joy_1	relish
joy_2	entertain
joy_2	excite
joy_2	enjoy that
joy_2	impress
joy_2	satisfy
joy_2	delight

joy_2	be mad about
joy_2	take pleasure in
joy_2	enjoy
joy_2	arouse
joy_2	be happy about
joy_2	please
joy_2	be lucky that
joy_2	cheer
joy_2	soothe
joy_2	captivate
joy_2	be proud of
joy_2	be satisfy that
joy_2	amuse
joy_2	savor
joy_3	calm
sadness	devastate
sadness	agonize over
sadness	regret that
sadness	pine for
sadness	worry about
sadness	be sad that
sadness	depress
sadness	regret
sadness_1	mourn that
sadness_1	grieve that
sadness_1	mourn
sadness_2	be unhappy that
sadness_2	upset
sadness_2	displease
sadness_2	be disappoint that
sadness_2	be unhappy about
sadness_2	be sadden that
sadness_2	disappoint
surprise	boggle
surprise	baffle
surprise	puzzle
surprise_1	be stun that
surprise_1	confound

surprise_1	stagger
surprise_1	stun
surprise_1	amaze
surprise_1	be bewilder that
surprise_1	be amaze that
surprise_1	wow
surprise_1	shock
surprise_2	surprise
surprise_2	startle
surprise_2	be flabbergast that
surprise_2	confuse
trust	charm
trust	be reassure that
trust_1	rely on
trust_2	trust
trust_2	take comfort in
trust_2	count on
trust_2	trust that
trust_2	reassure

Table 35: Patterns and their emotion/degree assigned by the majority

A.2 BIGRAMS

Bigram	PMI score
government reprisal	1.91
attack by:ethnic_albanian	1.90
possible reprisal	1.88
political backlash	1.87
long-term effect	1.87
reprisal attack	1.87
reprisal for:genocide	1.85
unintended consequence	1.85
new crackdown	1.84
government retribution	1.83

Table 36: Top 10 NP cause bigrams with highest fear PMI score

BIBLIOGRAPHY

[1] Charu C. Aggarwal and ChengXiang Zhai. *Mining text data*. Springer Science & Business Media, 2012.

[2] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics, 1998.

[3] Alexandra Balahur, Jesús M. Hermida, and Andrés Montoyo. Detecting implicit expressions of emotion in text: A comparative analysis. *Decision Support Systems*, 53(4):742–753, 2012.

[4] Jerome R. Bellegarda. Emotion analysis using latent affective folding and embedding. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 1–9. Association for Computational Linguistics, 2010.

[5] Steven Bird. NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics, 2006.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

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

[8] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 223–232. ACM, 2013.

[9] Dipankar Das and Sivaji Bandyopadhyay. Emotion holder for emotional verbs–the role of subject and syntax. In *Computational Linguistics and Intelligent Text Processing*, pages 385–393. Springer, 2010.

[10] Lingjia Deng, Yoonjung Choi, and Janyce Wiebe. Benefactive/-malefactive event and writer attitude annotation. In *ACL (2)*, pages 120–125, 2013.

[11] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[12] Paul Ekman and Wallace V. Friesen. *Unmasking the face: A guide to recognizing emotions from facial clues*. Ishk, 2003.

[13] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422. Citeseer, 2006.

[14] Christiane Fellbaum and Yvette Yannick Mathieu. A corpus-based construction of emotion verb scales. *Linguistic Approaches to Emotions in Context*, 241:99, 2014.

[15] J.R. Firth. A synopsis of linguistic theory, 1930–1955 selected papers of jr firth (1952–1959), fr palmer, 168 205, 1968.

[16] Matthew Gerber and Joyce Y. Chai. Semantic role labeling of implicit arguments for nominal predicates. *Computational Linguistics*, 38(4):755–798, 2012.

[17] Matthias Hartung and Anette Frank. Exploring supervised [lda] models for assigning attributes to adjective-noun phrases. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 540–551. Association for Computational Linguistics, 2011.

[18] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999.

[19] Karin Kipper-Schuler. *A Broad-Coverage Comprehensive Verb Lexicon*. PhD thesis, Computer and Information Science Dept., University of Pennsylvania, Philadelphia, PA, 2005.

[20] J. Richard Landis and Gary G. Koch. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, pages 363–374, 1977.

[21] Bing Liu. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, 2012.

[22] Cheng-Yu Lu, Shian-Hua Lin, Jen-Chang Liu, Samuel Cruz-Lara, and Jen-Shin Hong. Automatic event-level textual emotion sensing using mutual action histogram between entities. *Expert systems with applications*, 37(2):1643–1653, 2010.

[23] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 142–150. Association for Computational Linguistics, 2011.

[24] McCallum A. Mallet. A machine learning for language toolkit, 2002.

[25] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113, 2014.

[26] George A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[27] Saif Mohammad. From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 105–114. Association for Computational Linguistics, 2011.

[28] Saif M. Mohammad. # emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 246–255. Association for Computational Linguistics, 2012.

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

[30] Saif M. Mohammad and Tony Wenda Yang. Tracking sentiment in mail: how genders differ on emotional axes. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (ACL-HLT 2011*, pages 70–79, 2011.

[31] Saif M. Mohammad, Xiaodan Zhu, and Joel Martin. Semantic role labeling of emotions in tweets. 2013.

[32] Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pages 95–100. Association for Computational Linguistics, 2012.

[33] Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using em. *Machine learning*, 39(2-3):103–134, 2000.

[34] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

[35] Patrick Pantel and Marco Pennacchiotti. Automatically harvesting and ontologizing semantic relations. *Ontology learning and population: Bridging the gap between text and knowledge*, pages 171–198, 2008.

[36] Patrick Pantel and Deepak Ravichandran. Automatically labeling semantic classes. In *HLT-NAACL*, volume 4, pages 321–328, 2004.

[37] Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. English gigaword fifth edition. 2011.

[38] Robert Plutchik. A general psychoevolutionary theory of emotion. *Theories of emotion*, 1, 1980.

[39] Ashequl Qadir and Ellen Riloff. Learning emotion indicators from tweets: Hashtags, hashtag patterns, and phrases.

[40] Niklas Ravaja, Timo Saari, Marko Turpeinen, Jari Laarni, Mikko Salminen, and Matias Kivikangas. Spatial presence and emotions during video game playing: Does it matter with whom you play? *Presence: Teleoperators and Virtual Environments*, 15(4): 381–392, 2006.

[41] Alan Ritter, Oren Etzioni, et al. A latent dirichlet allocation method for selectional preferences. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 424–434. Association for Computational Linguistics, 2010.

[42] Thomas Scholz and Stefan Conrad. Opinion mining in newspaper articles by entropy-based word connections. In *EMNLP*, pages 1828–1839, 2013.

[43] Diarmuid O. Séaghdha. Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 435–444. Association for Computational Linguistics, 2010.

[44] Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642. Citeseer, 2013.

[45] Jacopo Staiano and Marco Guerini. DepecheMood: a lexicon for emotion analysis from crowd-annotated news. *arXiv preprint arXiv:1405.1605*, 2014.

[46] Philip J. Stone and Earl B. Hunt. A computer approach to content analysis: studies using the [g]eneral [i]nquirer system. In *Proceedings of the May 21-23, 1963, spring joint computer conference*, pages 241–256. ACM, 1963.

[47] Carlo Strapparava and Alessandro Valitutti. WordNet Affect: an affective extension of WordNet. In *LREC*, volume 4, pages 1083–1086, 2004.

[48] Yulia Tsvetkov, Nathan Schneider, Dirk Hovy, Archna Bhatia, Manaal Faruqui, and Chris Dyer. Augmenting English adjective senses with supersenses. In *Proc. of LREC*, 2014.

[49] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics, 2005.