Expert Systems with Applications 42 (2015) 7375–7385

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

The role of idioms in sentiment analysis

Lowri Williams [a], Christian Bannister [a], Michael Arribas-Ayllon [b], Alun Preece [a], Irena Spasić [a,*]

[a] School of Computer Science & Informatics, Cardiff University, 5 The Parade, Cardiff CF24 3AA, UK
[b] School of Social Sciences, Cardiff University, King Edward VII Avenue, Cardiff CF10 3WT, UK

ARTICLE INFO

Article history:
Available online 28 May 2015

Keywords:
Emotion recognition
Sentiment analysis
Natural language processing
User-generated content
Tagging

ABSTRACT

In this paper we investigate the role of idioms in automated approaches to sentiment analysis. To estimate the degree to which the inclusion of idioms as features may potentially improve the results of traditional sentiment analysis, we compared our results to two such methods. First, to support idioms as features we collected a set of 580 idioms that are relevant to sentiment analysis, i.e. the ones that can be mapped to an emotion. These mappings were then obtained using a web-based crowdsourcing approach. The quality of the crowdsourced information is demonstrated with high agreement among five independent annotators calculated using Krippendorff's alpha coefficient ($\alpha = 0.662$). Second, to evaluate the results of sentiment analysis, we assembled a corpus of sentences in which idioms are used in context. Each sentence was annotated with an emotion, which formed the basis for the gold standard used for the comparison against two baseline methods. The performance was evaluated in terms of three measures – precision, recall and *F*-measure. Overall, our approach achieved 64% and 61% for these three measures in two experiments improving the baseline results by 20 and 15 percent points respectively. *F*-measure was significantly improved over all three sentiment polarity classes: Positive, Negative and Other. Most notable improvement was recorded in classification of positive sentiments, where recall was improved by 45 percent points in both experiments without compromising the precision. The statistical significance of these improvements was confirmed by McNemar's test.

1. Introduction

The proliferation of user-generated content (e.g. product reviews) on the Web 2.0 provides opportunities for many practical applications that require consumer opinion (e.g. market research) as an alternative or a supplement to more traditional qualitative research methods such as surveys, interviews and focus groups. However, the sheer scale of text data acquired from the Web poses challenges to qualitative analysis. Text mining has emerged as a potential solution to the problems of information overload associated with reading vast amounts of text originating from diverse sources. In particular, sentiment analysis (or opinion mining) aims to automatically extract and classify sentiments (the subjective part of an opinion) and/or emotions (the projections or display of a feeling) expressed in text (Liu, 2010; Munezero, Montero, Sutinen, & Pajunen, 2014). Most research activities in this domain have focused on the problem of sentiment classification, which classifies an opinionated text segment (e.g. phrase, sentence or paragraph) in terms of its polarity: positive, negative or neutral (e.g. Aue & Gamon, 2005; Bethard, Yu, Thornton, Hatzivassiloglou, & Jurafsky, 2004; Breck, Choi, & Cardie, 2007).

Features used to support sentiment analysis include terms, part of speech, syntactic dependencies and negation (Pang & Lee, 2008). Most commonly, opinionated words that carry subjective bias are used in a bag-of-words approach to classify opinions (e.g. Attardi & Simi, 2006). Opinionated words can be utilized from lexicons such as SentiWordNet (Esuli & Sebastiani, 2006), WordNet-Affect (Valitutti, Strapparava, & Stock, 2004) and NRC word–emotion association lexicon (Mohammad & Turney, 2010). Dynamic calculation of word polarity (or semantic orientation) based on its statistical association with a set of positive and negative paradigm words is an alternative to predefined lexicons of opinionated words (Turney & Littman, 2003). Other features explored in sentiment analysis include more complex linguistic models based on lexical substitution, *n*-grams and phrases (Dave, Lawrence, & Pennock, 2003). Using an *n*-gram graph based method to assign sentiment polarity to individual word senses, experiments implied that figurative language (i.e. the language which digresses from literal meanings) not only conveys sentiment, but actually drives the polarity of a sentence (Rentoumi, Vouros, Karkaletsis, & Moser, 2012). Although the value of phrase-level features in sentiment

* Corresponding author. Tel.: +44 29 2087 0320; fax: +44 29 2087 4598.
E-mail addresses: WilliamsL10@cardiff.ac.uk (L. Williams), BannisterCA@cardiff.ac.uk (C. Bannister), Arribas-AyllonM@cardiff.ac.uk (M. Arribas-Ayllon), PreeceAD@cardiff.ac.uk (A. Preece), i.spasic@cs.cardiff.ac.uk (I. Spasić).

http://dx.doi.org/10.1016/j.eswa.2015.05.039
0957-4174/© 2015 The Authors. Published by Elsevier Ltd.

analysis has been acknowledged (Socher et al., 2013; Wilson, Wiebe, & Hoffmann, 2009), few approaches have extensively explored idioms as features of this kind (e.g. Thelwall, Buckley, & Paltoglou, 2012). Nonetheless, the error analysis of sentiment classification results often reveals that the largest percentage of errors are neutral classifications when no opinionated words are present or when idioms are used to express sentiment (Balahur et al., 2010).

Idioms are often defined as multi-word expressions, the meaning of which cannot be deduced from the literal meaning of constituent words, e.g. the idiom *a fish out of water* is used to refer to someone who feels uncomfortable in a particular situation. To distinguish idioms from related linguistic categories such as formulae, fixed phrases, collocations, clichés, sayings, proverbs and allusions, the following properties need to be considered (Nunberg, Sag, & Wasow, 1994):

1. *Conventionality*: Their meaning cannot be (entirely) predicted from the constituent words considered independently.
2. *Inflexibility*: Their syntax is restricted, i.e. idioms do not vary much in way they are composed.
3. *Figuration*: Idioms typically have figurative meaning stemming from metaphors, hyperboles and other types of figuration.
4. *Proverbiality*: Idioms usually describe a recurrent social situation.
5. *Informality*: Idioms are associated with less formal language such as colloquialism.
6. *Affect*: Idioms typically imply an affective stance toward something rather than a neutral one.

The last property emphasizes the importance of idioms in sentiment analysis as it implies that an idiom itself may often be sufficient to determine the underlying sentiment. There are two requirements for idioms to be effectively utilized in sentiment analysis methods: (1) Idioms need to be recognized in text, and (2) the associated sentiment needs to be explicitly encoded.

The inflexibility property (see property 2 above) makes the first requirement feasible. Lexico-syntactic patterns can be used to model idioms computationally and recognize their occurrences in text. A lot of the idioms are frozen phrases such as *by and large*, which can be recognized by simple string matching. Syntactic changes such as inflection (e.g. verb tense change) are often seen in idioms (Yusifova, 2013). Such linguistic phenomena can be modeled by regular expressions, e.g. *spill[s|t|ed] the beans*. More complex idioms have variables for open argument places (Jackendoff & Pinker, 2005) (e.g. *put someone in one's place*), which can still be modeled by means of lexico-syntactic patterns (e.g. *put NP in PRN's place*) and recognized in a linguistically pre-processed text. Less often, idioms are "syntactically productive", i.e. they can be changed syntactically without losing their figurative meaning, e.g. *John laid down the law* can be passivized to *the law was laid down by John* while retaining the original figurative interpretation that John enforced the rules (Gibbs & Nayak, 1989). Transformational grammars have been suggested as a framework to handle more complex syntactic changes such as nominalization (e.g. *you blew some steam off* vs. *your blowing off some steam*) (Fraser, 1970).

In Polish, a highly inflected language, idioms were recognized using a cascade of regular expressions and their effect on sentiment analysis results was evaluated on a corpus of product and service reviews, where idioms were found to occur rarely (Buczynski & Wawer, 2008). In English, despite the obvious need for regular expressions, idioms are usually recognized using a lexicon-based approach, which can only recognize those idioms that are syntactically unproductive or frozen. For example, (Shastri, Parvathy, Abhishek, J., & R., 2010) used a dictionary of idioms (e.g. *at a snail's pace*) in order to recognize them in text and map them to their abstract meaning (e.g. *slow*), which is then utilized to infer the sentiment. In another lexicon-based approach (Beigman Klebanov, Burstein, & Madnani, 2013), idiom recognition was further limited to 46 noun–noun compounds (e.g. *glass ceiling*). The use of their sentiment profiles was found to improve the performance of sentiment classification on a corpus of test-takers essays. Our own study aims to go beyond a lexicon-based approach to recognition of English idioms and use regular expressions instead. The added overhead of handcrafting regular expressions allowed us to explore a much wider set of idioms (beyond the *low-hanging fruits*) as part of sentiment analysis.

Assuming that idioms can be identified in text automatically, we need additional knowledge about the underlying sentiment in order to utilize them as features of sentiment analysis. While idioms have been extensively studied across many disciplines (e.g. linguistics, psychology, etc.), thus far there is no comprehensive knowledge base that systematically maps idioms to sentiments. This is the main reason why idioms have been underrepresented as features used in sentiment analysis approaches with few exceptions (e.g. Xie and Wang (2014) describe a set of 8160 Chinese idioms). Due to the subjective nature of the problem, multiple annotations are required in order to either determine the prevalent sentiment associated with an idiom or use a fuzzy logic approach to represent the sentiment with a degree of truthfulness and falsehood. To support this task, a web-based crowdsourcing approach can be used to efficiently collect a large amount of information relevant for sentiment analysis (Greenwood, Elwyn, Francis, Preece, & Spasić, 2013). We used crowdsourcing to systematically map 580 English idioms to 10 emotion categories, which represents the largest lexico-semantic resource of this kind to utilize in sentiment analysis.

The purpose of this paper is to study the effect of idioms on sentiment analysis. The study was designed as follows (see Fig. 1): (1) collect a set of idioms that can be mapped to sentiments, (2) map individual idioms to sentiments in order to support them as features for sentiment analysis, (3) assemble a corpus of sentences in which these idioms are used, (4) annotate sentences in the corpus with sentiments in order to create a gold standard for sentiment analysis, (5) implement a sentiment analysis approach that incorporates idioms as features, and (6) compare the evaluation results against a traditional sentiment analysis approach using the gold standard created in (4).

2. Data collection

2.1. Idioms

Idioms pose considerable difficulties for English language learners. Failure to understand idioms in context significantly affects one's understanding of language in a variety of personal and professional situations (Nippold & Martin, 1989). It is therefore not surprising that most syllabi for English as a second language pay special attention to studying idioms (Liu, 2003). As a result, there is an abundance of teaching material dedicated to the study of idioms. In this study, we relied upon an educational web site – *Learn English Today* (Learn English Today, 2013), which organizes idioms by themes, many of which can be mapped to emotions either directly (e.g. *Happiness/Sadness*) or indirectly (e.g. *Success/Failure*). We focused specifically on emotion-related idioms, as these are anticipated to have a substantial impact on sentiment analysis. We selected 16 out of a total of 60 available themes, listed in Table 1 together with a number of associated idioms. A total of 580 idioms were collected.

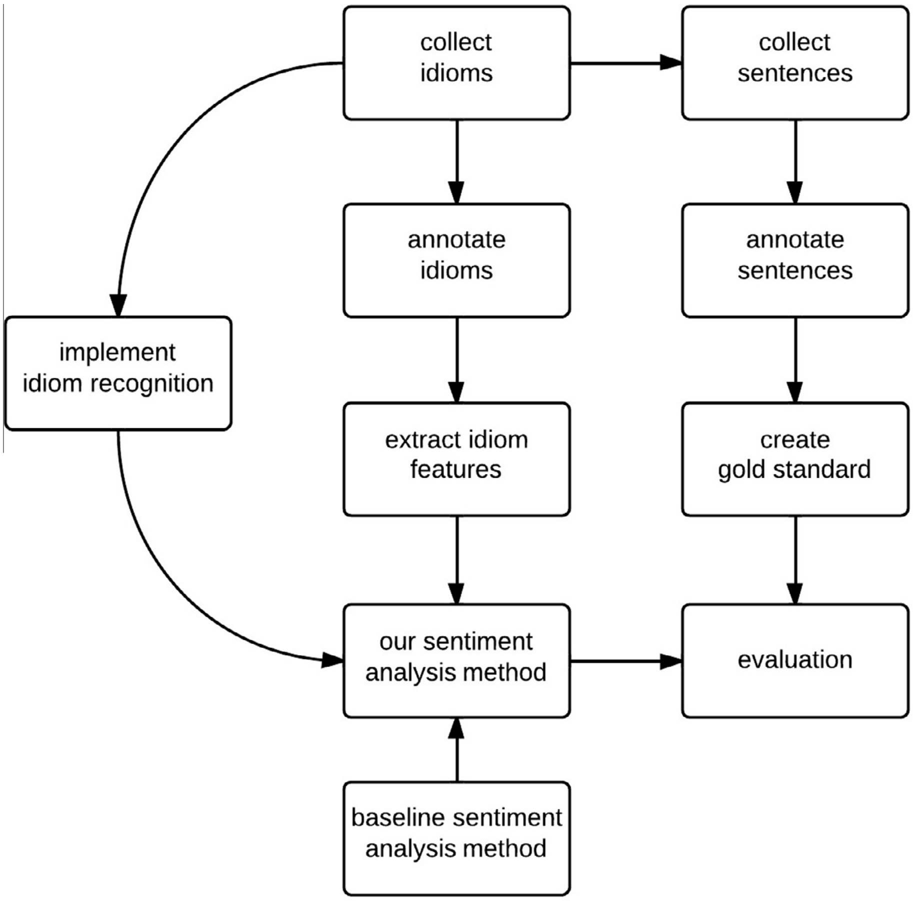


Fig. 1. An overview of the study design.

Table 1
Distribution of idioms across themes.

Theme	Total	Theme	Total
Anger/Annoyance	45	Mistakes/Errors	5
Anxiety/Fear	14	Politeness	8
Arguments/Disagreements	37	Problems/Difficulties	57
Enthusiasm/Motivation	10	Safety/Danger	27
Feelings/Emotions	48	Sleep/Tiredness	11
Fun/Enjoyment	22	Success/Failure	84
Happiness/Sadness	21	Surprise/Disbelief	16
Madness/Insanity	11	Violence	6

2.2. Corpus

The British National Corpus (BNC) (BNC Consortium., 2014; Leech, 1993) is a large text corpus of both written and spoken English compiled from a variety of sources. As such, it has been a corpus of choice for a great many studies in computational linguistics and natural language processing (NLP), including those focused on idioms (e.g. Grant, 2005). We also used the BNC to assemble a corpus of idioms used in context. We collected 2521 sentences that contain an expression that can be matched to an idiom. In most cases, this expression will have a figurative meaning associated with an idiom, but in some cases it will convey a literal meaning. In this sense, some of the sentences will be false positives. From a lexico-syntactic perspective, most idioms can be modeled with local grammars, but it is more difficult to automate their recognition from a semantic perspective. Consider, for example, the following two sentences extracted for expression *in the bag*, the figurative meaning of which is *virtually secured*:

The Welsh farmer's son had the 1988 conditional jockeys' title already in the bag.
I looked in the bag, it was full of fish.

It is necessary to include false positives in the corpus in order to evaluate how incorrectly recognized idioms may affect the results of sentiment analysis.

The BNC was searched using its Simple Search function available online. It can be used to search the BNC for a word of phrase and returns up to 50 random sentences for each query. The BNC was searched for content words found in idioms and the returned results were manually matched to an idiom. A maximum of 10 sentences was selected for each idiom used in either a figurative or literal sense. Search results were non-empty for a total of 423 idioms from the original list of 580 idioms. The mean and median average number of sentences extracted for an idiom are both 6, with standard deviation of 3.39. Table 2 summarizes the number of sentences collected for each theme associated with idioms.

Table 2
Distribution of sentences across themes.

Theme	Total	Theme	Total
Anger/Annoyance	261	Mistakes/Errors	31
Anxiety/Fear	88	Politeness	42
Arguments/Disagreements	232	Problems/Difficulties	360
Enthusiasm/Motivation	41	Safety/Danger	176
Feelings/Emotions	280	Sleep/Tiredness	64
Fun/Enjoyment	107	Success/Failure	519
Happiness/Sadness	128	Surprise/Disbelief	92
Madness/Insanity	47	Violence	50

3. Data annotation

3.1. Annotation scheme

Taking a long-term view of our research on sentiment analysis, we want to address the limitations of state-of-the-art approaches by focusing on a full range of emotions rather than merely sentiment polarity. However, there is no consensus among researchers about how to define, measure or conceptually organize the range of feelings and behaviors that correspond to an emotion. The main tension in the literature is whether emotions can be defined as discrete, universal categories of basic emotions or whether they are characterized by one or more dimensions. Categorical approaches are usually theory-driven accounts that suggest basic emotions are the functional expression of underlying biological and evolutionary processes (Damasio, 1999; Darwin, 1872; LeDoux, 1996). This view is supported by empirical findings of cross-cultural studies based on recognition of facial expressions (Ekman, 1972). Ekman claimed that there are six basic emotions – anger, disgust, fear, happiness, sadness and surprise – though later studies identified additional emotions (Ekman, 1992). Dimensional approaches are often data driven accounts of emotions incorporating aspects of valence, arousal or intensity derived from empirical studies of self-reported experience (Russell, 2003). There is considerable variation among dimensional models, many of which incorporate either two or three dimensions (Rubin & Talarico, 2009).

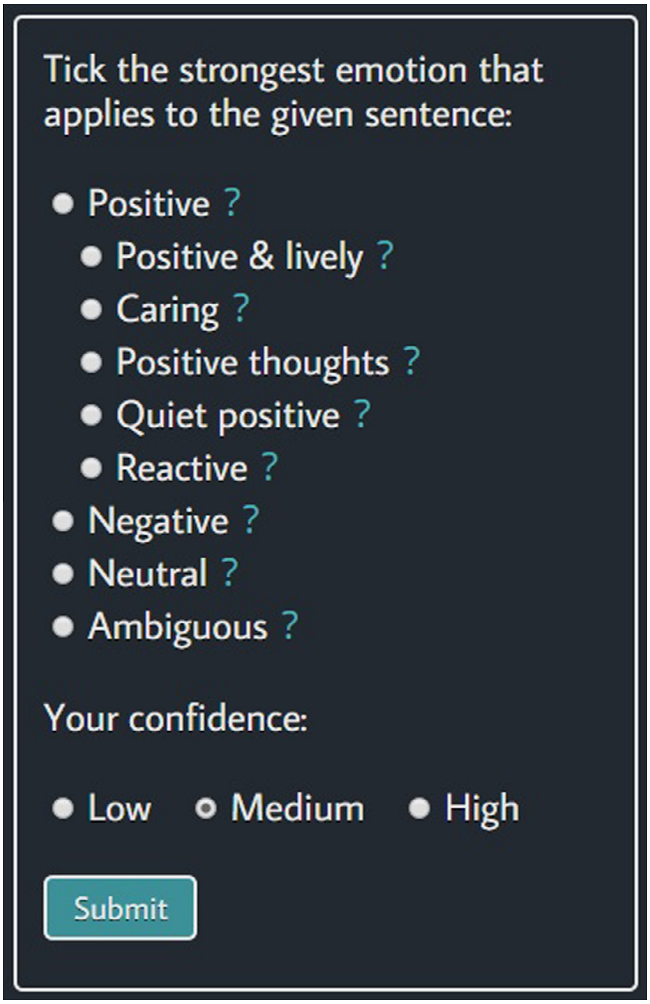
Having considered the literature on emotion classification, traditional taxonomies arising from psychological research are too complex for our present task. We therefore opted to base our annotation scheme on the Emotion Annotation and Representation Language (EARL) Version 0.4.0, an XML-based language for representing and annotating emotions in technological contexts (The Association for the Advancement of Affective Computing, 2014). It has been designed for a wide range of applications, including corpus annotation and emotion recognition. It organizes 48 emotions into 5 positive and 5 negative categories (see Table 3). To facilitate the annotation task, we used the 10 top-level categories as they provide a manageable number of choices for a human annotator, which are also evenly distributed between positive and negative polarities. For annotation purposes, specific emotions were used as examples instead of formal definitions to explain each top-level category, e.g. *Caring* encompasses *affection, empathy, friendliness* and *love*.

3.2. Annotation process

We implemented a bespoke web-based annotation platform. It provides a simple interface accessible via a web browser, which eliminates the installation overhead and minimizes the need for special training. The interface consists of two panes. One pane contains randomly selected text ready to be annotated. In our study, this was either an idiom together with its definition or a sentence that contains an idiom (see Fig. 1 for study design). The other pane contains four annotation choices: *Positive, Negative, Neutral* and *Ambiguous*. The selection of either *Positive* or *Negative* categories expands the menu to provide additional choices (see Fig. 2). The category *Neutral* was introduced in EARL to allow the absence of an emotion to be annotated explicitly. We introduced the additional category *Ambiguous* to annotate cases of emotion that cannot be determined as either *Positive* or *Negative* without additional information such as context, tone of voice, or body language.

Table 3
Top-level emotion categories in EARL.

Negative category	Example	Positive category	Example
Negative & forceful	Annoyance	Positive & lively	Joy
Negative & not in control	Helplessness	Caring	Love
Negative thoughts	Doubt	Positive thoughts	Hope
Negative & passive	Sadness	Quiet positive	Relaxed
Agitation	Stress	Reactive	Politeness



Tick the strongest emotion that applies to the given sentence:

- ● Positive ?
  - ● Positive & lively ?
  - ● Caring ?
  - ● Positive thoughts ?
  - ● Quiet positive ?
  - ● Reactive ?
- ● Negative ?
- ● Neutral ?
- ● Ambiguous ?

Your confidence:

● Low ◉ Medium ● High

Fig. 2. Annotation pane.

A help button is available next to each annotation choice to provide additional information if needed. The overall annotation scheme can also be viewed in a separate window. Each annotation can be scored on a 3-point scale to account for the annotator's confidence in a particular choice: *Low, Medium* or *High*, with *Medium* being a default choice.

Online accessibility provided us with the flexibility of choosing the physical location for annotation experiments. Users were tracked by their IP addresses to avoid duplication of annotations, not to identify individuals. No other personal information was collected. This was explained in the privacy policy. All annotation results were stored securely in a relational database.

Group annotation sessions were conducted weekly, where new annotators were briefed about the study and their role as an annotator. All annotators were required to be of native or native-like English proficiency. All annotations were performed independently. The data were randomized individually for each annotator, so they were always annotated in a different order. No discussions about particular data items were allowed among the annotators during the group sessions.

3.3. Annotation results

The data was annotated over a course of 13 weeks. A total of 18 annotators participated in the study with 44% being female. A total of 2900 annotations were collected for all 580 idioms described in Section 2.1 with 5 annotations per idiom. A total of 8610 annotations were collected for all sentences in the corpus described in Section 2.2 with at least 3 annotations per sentence. A total of 143 sentences had a maximum of 5 annotations. Overall, the mean and median average number of annotations per sentence were both 3 with a standard deviation of 0.60.

The main goal of this study was to evaluate the role of idioms in sentiment analysis by comparing our method to existing approaches, hence our experiments needed to conform to their sentiment classification scheme. Most sentiment analysis approaches output sentiment polarity, i.e. classify text as being positive, negative or neutral. This also applies to SentiStrength (Thelwall, 2014; Thelwall, Buckley, Paltoglou, Cai, & Kappas, 2010) and Stanford CoreNLP's sentiment annotator (Socher et al., 2013), two state-of-the-art methods chosen as the baseline in our experiments. Therefore, to create a gold standard that can allow comparison against the baseline, we projected categories specified in Table 3 onto *Positive* and *Negative* polarity and merged *Neutral* and *Ambiguous* categories into a single category called *Other*. Nonetheless, we will be able to use the original annotations to re-train the machine learning method described here to support emotion classification against the categories described in Section 3.1 as part of the future work.

After projecting all annotations onto the simplified classification scheme (i.e. *Positive*, *Negative* and *Other*), we used Krippendorff's alpha coefficient (Krippendorff, 1980) to measure the inter-annotator agreement. As a generalization of known reliability indices, it was chosen because it applies to: (1) any number of annotators, not just two, (2) any number of categories, and (3) incomplete or missing data (Krippendorff, 2004). Krippendorff's alpha coefficient of 1 indicates perfect agreement, whereas 0 indicates chance agreement. Therefore, higher values indicate better agreement.

The values for Krippendorff's alpha coefficient were obtained using an online tool for calculating inter-annotator agreement (Geertzen, 2012). The expected disagreement on the idiom dataset was calculated to be $D_e = 0.606$, the observed disagreement was $D_o = 0.205$, which resulted in $\alpha = 0.662$. The corresponding values calculated for the corpus of sentences were as follows: $D_o = 0.414$, $D_e = 0.643$ and $\alpha = 0.355$. The relatively high agreement ($\alpha = 0.662$) on idioms alone illustrates that idioms can be reliably mapped to sentiment polarity. Significantly lower agreement ($\alpha = 0.355$) on sentences where idioms are used in context illustrates the complexity of sentiment interpretation, where a range of different emotions may be conveyed in a single sentence (e.g. *They were too embarrassed or didn't want to make a mountain out of a molehill*) or the underlying emotion may vary depending on the context (e.g. *His jaw dropped*).

3.4. Gold standard

Annotated sentences were used to create a gold standard for sentiment analysis experiments. Prior to calculating the inter-annotator agreement, additional annotations from a new independent annotator were sought to resolve any disagreements. The annotation process was finalized once each sentence had the majority of at least 50% across the annotations with all previous ties broken. A total of 282 additional annotations were collected in three iterations. Subsequently, each sentence with an annotation agreed by the relative majority of at least 50% of the annotators was treated as the ground truth.

Table 4 shows the distribution of ground truth annotations across the three categories together with an annotated example from each. A random subset of approximately 10% was selected for testing (500 sentences) and the rest was used for training (2021 sentences).

4. Sentiment analysis

4.1. Idiom recognition

In order to incorporate idioms as features of sentiment analysis, we require the means of recognizing them in text. For this purpose, we modeled each idiom by a local grammar (Gross, 1997). For example, the following grammar:

```
⟨idiom⟩ ::= ⟨VB⟩ ⟨PRP$⟩ heart on ⟨PRP$⟩ sleeve
⟨VB⟩ ::= wear | wore | worn | wearing
⟨PRP$⟩ ::= my, your, his, her, its, our, their
```

was used to successfully recognize the idiom *wear one's heart on one's sleeve* in the following sentence:

Rather than ⟨idiom⟩ *wear your heart on your*
sleeve ⟨/idiom⟩, *you keep it under your hat.*

Idiom recognition rules were implemented as expressions in Mixup (My Information eXtraction and Understanding Package), a simple pattern-matching language (Cohen, 2014). The pattern-matching rules were applied to the test dataset of 500 sentences in which a single annotator marked up all idiom occurrences differentiating between the figurative and literal meaning, e.g.

Phew, that was a ⟨idiom⟩ *close shave* ⟨/idiom⟩.
He has polished shoes, a ⟨nonidiom⟩ *close shave* ⟨/nonidiom⟩, *and too much pride for a free drink.*

The following performance was recorded for idiom recognition: $P = 94.44\%$, $R = 100\%$ and $F = 97.14\%$, where an idiom was considered to be correctly recognized if the suggested text span matched exactly the one marked up by the annotator.

4.2. Negation

As with other phrases, the polarity of idioms can be changed by negation. For example, the polarity of the idiom *jump for joy* is positive, but when negated the overall polarity can be reversed as illustrated by the following sentence:

I didn't exactly jump for joy.

It is, therefore, essential to consider negation when using idioms as features in sentiment analysis. We implemented pattern-matching rules to recognize explicitly negated idioms based on clues such as negative words (e.g. *no, never*, etc.), negative adverbs (e.g. *hardly, barely*, etc.) and negated verbs (e.g. *doesn't, isn't*, etc.). A total of 27 negated idioms were annotated in the test

Table 4
Distribution of annotations in the gold standard.

Annotation	Total	%	Example
Negative	1219	48.35	*All right, do not jump down my throat*
Positive	677	26.85	*I shall go the extra mile*
Other	625	24.79	*Your mother used to sleep like a log*

dataset. The following performance was recorded for the recognition of negated idioms: $P = 86.21\%$, $R = 92.59\%$ and $F = 89.29\%$. Given a small number of negated idioms in the test dataset, a larger corpus is required in order to better estimate the performance of the negation module.

4.3. Feature selection

Once idioms are recognized in text, we require additional information about their sentiment polarity in order to utilize them as features of sentiment analysis. Note that the polarity information was obtained as part of data annotation (see Section 3) when each idiom was annotated independently 5 times. After projecting the original annotations onto the sentiment polarity scheme, a total of 5 annotations collected for each idiom were used to calculate their feature vectors. Each idiom was represented by a triple (*Positive, Negative, Other*), where each value represents the percentage of annotations in the corresponding category. For example, the idiom *wear one's heart on one's sleeve* received one *Positive* annotation, zero *Negative* annotations and four *Other* annotations. It was, therefore, represented as the following triple: (20, 0, 80).

Further, as we wanted to investigate the impact of idioms as features in sentiment analysis, we conducted two experiments in which we combined idioms with the results of two popular sentiment analysis methods: (1) SentiStrength (Thelwall, 2014; Thelwall et al., 2010) and (2) Stanford CoreNLP's sentiment annotator (Socher et al., 2013).

In Experiment 1, we used SentiStrength as the baseline method and used its output as features in our own method and combined them with those based on idioms. SentiStrength is a rule-based system that assigns sentiment polarity to a sentence by aggregating the polarity of individual words, e.g.

Input: *The party is over.*
Analysis: The party [1] is over [−1] .
Output: result = 0, positive = 1, negative = −1

As illustrated in the given example, we used trinary classification output as features in our method and converted them into a 3-dimensional vector: (0, 1, −1). In our approach, the phrase *party is over* would be recognized as an idiom, which was annotated and subsequently mapped to the following triple: (0, 100, 0) denoting that all annotators considered it to be negative. We appended the two vectors to create a single feature vector for the given sentence as follows:

$$(\underbrace{0, 1, -1}_{\substack{\text{sentiment}\\ \text{polarity}}}\ \underbrace{0, 100, 0}_{\substack{\text{idiom}\\ \text{polarity}}})$$

In Experiment 2, we used a sentiment annotator distributed as part of the Stanford CoreNLP, a suite of core NLP tools. This method uses recursive neural networks to perform sentiment analysis at all levels of compositionality across the parse tree by classifying each sub-tree on a 5-point scale: very negative, negative, neutral, positive and very positive (see Fig. 3 for an example). In addition to classification, it also provides probability distribution across the 5 classes, which we have used as features in our method by converting them into a 5-dimensional vector: (4, 27, 46, 20, 3). As before, the idiom *party is over* would be recognized and mapped to its polarity triple (0, 100, 0) and appended to create a single feature vector for the given sentence as follows:

$$(\underbrace{4, 27, 46, 20, 3}_{\substack{\text{sentiment}\\ \text{polarity}}}\ \underbrace{0, 100, 0}_{\substack{\text{idiom}\\ \text{polarity}}})$$

If an idiom was recognized to be negated, then we reversed positive and negative polarities in the idiom polarity vector based on an assumption that negation typically converts positive to negative polarity and vice versa. For instance, let us consider the following sentence in which the negatively charged idiom *party is over* is negated:

The party is not over yet.

In this particular example, the negation changes the overall polarity from negative to positive. We modeled this phenomenon associated with negation by simply reversing positive and negative polarities, e.g. the original idiom polarity triple (0, 100, 0) would be converted to (100, 0, 0).

When multiple idioms are recognized, then the associated idiom polarity values are aggregated by summing up the polarity vectors previously taking into account the effects of negation. For instance, two idioms were recognized in the following sentence:

He ⟨idiom⟩ *stopped dead in his tracks* ⟨/idiom⟩, ⟨idiom⟩ *rooted to the spot* ⟨/idiom⟩ *with horror.*

Their polarity triples (0, 60, 40) and (0, 40, 60) respectively are summed up to obtain (0, 100, 100), which illustrates how the negative polarity is boosted from 60 and 40 to 100 in this particular case. In our corpus of 2521 sentences, 29 sentences contained more than one idiom. Finally, if no idiom was detected, then the idiom polarity values are set to zero: (0, 0, 0).

4.4. Sentiment classification

Sentiment analysis can be viewed as a classification problem with three available classes: *Positive*, *Negative* and *Other*. A wide range of supervised learning approaches can be used for this purpose. We used Weka (Hall et al., 2009), a popular suite of machine learning software, to train a classifier and perform classification experiments. We based our choice of a machine learning method on the results of cross-validation experiments on the training dataset. We opted for a naïve Bayes classifier, more specifically a Bayesian network classifier, which outperformed other methods available in Weka. For example, a naïve Bayes classifier outperformed support vector machines in terms of *F*-measure, proved more robust against the choice of features and provided more balanced classification performance across different classes. These findings agree with previous experiments when we successfully used a naïve Bayes approach to classify sentences by emotions they convey (Spasić, Burnap, Greenwood, & Arribas-Ayllon, 2012) and can be partially explained by the fact that a naïve Bayes classifier does not necessarily require a lot of training data to perform well (Domingos & Pazzani, 1997).

5. Evaluation

To evaluate the impact of ignoring idioms in sentiment analysis, we conducted two experiments initially outlined in Section 4.3. In Experiment 1, we used SentiStrength (Thelwall, 2014; Thelwall et al., 2010) as the baseline method as well as a part of feature selection for our own method. In Experiment 2, we did the same with the Stanford CoreNLP's sentiment annotator (Socher et al., 2013). The classification performance was evaluated in terms of three measures – precision (*P*), recall (*R*) and *F*-measure based on the numbers of true positives (TP), false positives (FP) and false negatives (FN). Tables 5 and 6 provide the comparison of these values for the two methods considered. The overall performance represents micro-averaged results across the three classes.

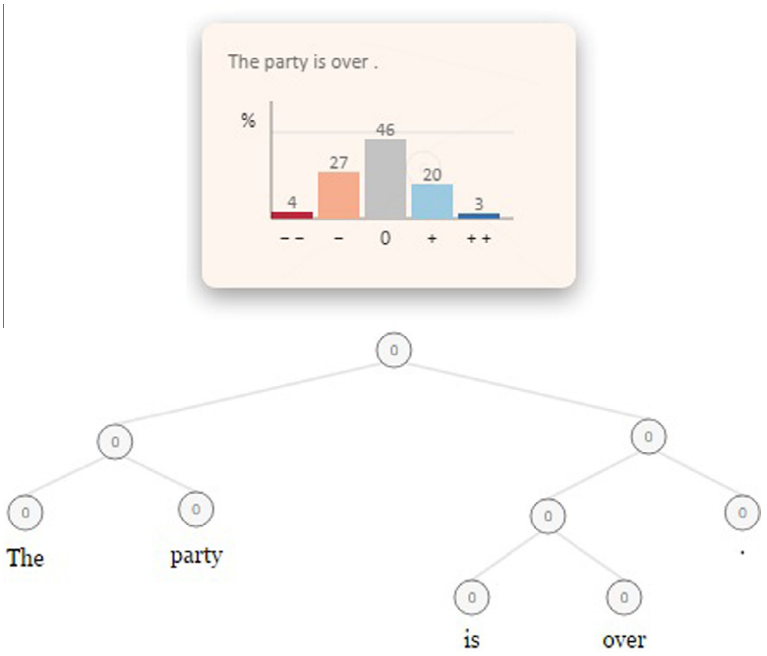


Fig. 3. Sentiment analysis results from Stanford CoreNLP.

Table 5
The evaluation results using SentiStrength as the baseline method.

Class	Method	TP	FP	FN	*P*	*R*	*F*
Positive	Baseline	40	53	98	43.01	28.99	34.63
	Ours	102	63	36	61.82	73.91	67.33
Negative	Baseline	111	66	127	62.71	46.64	53.49
	Ours	170	54	68	75.89	71.43	73.59
Other	Baseline	72	158	52	31.30	58.06	40.68
	Ours	49	62	75	44.14	39.52	41.70
Overall	Baseline	223	277	277	44.60	44.60	44.60
	Ours	321	179	179	64.20	64.20	64.20

With the exception of recall for *Other* class in Experiment 1 (a drop from 58.06% to 39.52%), our method demonstrates a considerable improvement across all three measures. The overall improvement in Experiment 1 is 19.60 percent points (see Table 5) and 15 percent points in Experiment 2 (see Table 6). In terms of *F*-measure, there is an improvement across all three classes, but most notably in *Positive* classifications. This is mainly due to improving the recall considerably by 45 percent points in both experiments without compromising the precision.

Confusion matrices given in Tables 7 and 8 show how classification outcomes are re-distributed across the three classes. Table 7a illustrates that SentiStrength is conservative in making both *Positive* and *Negative* predictions, thus less often misclassifying instances of *Other* class. Its classification outcomes were improved in all other cases by the use of idiom-based features (see Table 7b). Conversely, Stanford CoreNLP sentiment annotator proved to be more conservative in making *Positive* predictions compared to making *Negative* ones (see Table 8a), thus making fewer misclassifications when making *Positive* predictions. Nonetheless, its classification outcomes were improved in all other cases by the use of idiom-based features (see Table 8b).

Table 6
The evaluation results using Stanford CoreNLP sentiment annotator as the baseline method.

Class	Method	TP	FP	FN	*P*	*R*	*F*
Positive	Baseline	41	44	97	48.24	29.71	36.77
	Ours	104	81	34	56.22	75.36	64.40
Negative	Baseline	170	160	68	51.52	71.43	59.86
	Ours	181	82	57	68.82	76.05	72.26
Other	Baseline	19	66	105	22.35	15.32	18.18
	Ours	20	32	104	38.46	16.13	22.73
Overall	Baseline	230	270	270	46.00	46.00	46.00
	Ours	305	195	195	61.00	61.00	61.00

Table 7
Confusion matrices using SentiStrength as the baseline method.

Actual	Predicted		
	P	N	O
(a) Baseline			
P	40	36	62
N	31	111	96
O	22	30	72
(b) Our method			
P	102	18	18
N	24	170	44
O	39	36	49

Table 8
Confusion matrices using Stanford CoreNLP sentiment annotator as the baseline method.

Actual	Predicted		
	P	N	O
(a) Baseline			
P	41	74	23
N	25	170	43
O	19	86	19
(b) Our method			
P	104	24	10
N	35	181	22
O	46	58	20

Finally, in order to determine the statistical significance of the improvement over the baseline methods we performed the analysis of paired observations. We compared the sentiment classification results for each sentence before and after taking idioms in consideration by using continuity corrected McNemar's test (Everitt, 1977) to check for statistically significant differences in error rates. Under the null hypothesis, the two methods compared should have the same error rate.

McNemar's test is based on the χ^2 test statistic and (approximately) distributed as χ^2 with 1 degree of freedom. We used a variant of McNemar's test statistic that incorporates a correction for continuity to account for the fact that the statistic is discrete while the χ^2 distribution is continuous. The choice of this particular test was based on the following two facts: (1) McNemar's test has been shown to have low type I error, in this case – the probability that it would incorrectly detect a difference when no difference exists. (2) Its statistical power is improved when compared with the commonly used paired *t*-test (Dietterich, 1998).

The specific $\chi^2(1)$ and *p*-values recorded for the data produced in Experiment 1, where SentiStrength was used as the baseline method, were $\chi^2(1) = 43.16$ and $p < 0.001$. The values recorded for Experiment 2, where a sentiment annotator distributed as part of the Stanford CoreNLP was used as the baseline method, were $\chi^2(1) = 29.28$ and $p < 0.001$. Therefore, in both cases the results of McNemar's test confirmed that there was a statistically significant difference in error rates between the two methods.

6. Discussion

We evaluated the impact of idioms as features of sentiment analysis by showing that they significantly improve classification results when such features are present. Their overall impact on sentiment classification can be estimated by combining this information with idiom distribution. For this purpose, we used corpora commonly used to evaluate performance of sentiment analysis (see Table 9). We used the idiom recognition module described in Section 4.1 to match a list of 580 idioms against the given corpora. Table 10 provides the number of matched idiom occurrences, the number of different idioms matched and the ratio of idiom occurrences against the corpus size in megabytes. These values illustrate that the distribution of idioms varies considerably across different genres.

Idioms were most commonly found in the movie reviews (MR1–MR3). Focusing on full-text reviews, we observed that 6.02% of documents from corpus MR1 contained idioms, whereas this number in corpus MR2 rose to 18.95%, thus illustrating significant impact idioms may have on document classification. If we compare full-text reviews from corpus MR2 to subjective snippets of such reviews from corpus MR3, we can conclude that subjective sentences are more likely to contain idioms than the rest of the text, again suggesting the value of idioms in sentence classification in terms of their sentiment polarity.

Table 10
Distribution of idioms across the corpora.

Corpus	Idioms	Unique idioms	Ratio
MR1	3250	359	51.26
MR2	464	161	62.53
MR3	75	44	63.56
HR	3904	294	17.05
CR	393	101	14.14
PR	462	113	36.09
TW	2788	309	36.54

As illustrated by corpora HR and CR, idioms were less commonly found in hotel and car reviews. Contrary to previous findings that short conversations normally contain fewer idioms (Straessler, 1982), tweets proved to contain a relatively high proportion of idioms, which, at ratio of 36.54, is in line with that of product reviews.

The number of unique idioms found was related to corpus size. Naturally, the highest number of different idioms was found in the largest corpora: MR1, HR and TW. More importantly, the variety of idioms used was found to be strongly correlated to the genre rather than the size. The use of idioms in each corpus followed the power law distribution with a small number of idioms used frequently and the rest used rarely, but the frequently used idioms differed across the genres. To explore the bias in idiom usage, we selected top 20 most frequently occurring idioms in each corpus and compared these sets using the Jaccard similarity coefficient (see Table 11), which is defined as the size of the intersection divided by the size of the union of the given sets. These similarities were used to cluster corpora based on idiom usage. Fig. 4 provides a dendrogram produced as a result of hierarchical clustering based on complete linkage and Euclidian distance, whose values are shown between the clusters. Table 12 illustrates the differences in most frequently used idioms across the corpora.

Our attempt to generalize the evaluation results presented in Section 5 will be based on the following assumptions. Given a corpus of subjective sentences, let *I* refer to a subset of such sentences that contain idioms and let *O* refer to the remaining sentences. Further, let F_I and F_O denote the *F*-measure values achieved by the baseline sentiment classification method on these two subsets respectively. If we re-classify the sentiment on these subsets by combining the baseline method with idiom-based features as described in Section 4, then the value of F_O will remain unchanged, whereas the *F*-measure on set *I* is expected to increase to $F_I + i$, where *i* refers to the expected improvement. If *p* is the percentage of subjective sentences that contain idioms (i.e. $p = |I|/|I \cup O|$), then we can use it to roughly estimate the overall value of the *F*-measure as the weighted average of the values achieved on the given subsets, i.e. $F = p \cdot (F_I + i) + (1 - p) \cdot F_O$.

In that case, the overall improvement of the *F*-measure can be estimated as the value of the product $p \cdot i$. In other words, in order to generalize the evaluation results presented in Section 5, the improvement of sentiment classification on sentences with idioms

Table 9
Description of the corpora.

Identifier	Document type	Original source	Coverage	Size	Source
MR1	Movie reviews	http://www.imdb.com/	50,000 reviews	63.40 MB	(Maas et al., 2011)
MR2	Movie reviews	http://www.rottentomatoes.com/	2000 reviews	7.42 MB	(Pang & Lee, 2004)
MR3	Movie reviews	http://www.rottentomatoes.com/	10,662 snippets	1.18 MB	(Pang & Lee, 2005)
HR	Hotel reviews	http://www.tripadvisor.co.uk/	259,000 reviews	229.00 MB	(Ganesan, Zhai, & Han, 2010)
CR	Car reviews	http://www.edmunds.com/	42,230 reviews	27.80 MB	(Ganesan et al., 2010)
PR	Product reviews	http://www.cnet.com/	300 products	12.80 MB	(Ganesan, Zhai, & Viegas, 2012)
TW	Tweets	http://twitter.com/	1,048,576 tweets	76.30 MB	(Go, Bhayani, & Huang, 2009)

Table 11
Similarity of idiom usage across the corpora.

Corpus	MR1 (%)	MR2 (%)	MR3 (%)	HR (%)	CR (%)	PR (%)	TW (%)
MR1	100.00	33.33	25.00	11.11	8.11	8.11	17.65
MR2	33.33	100.00	17.65	14.29	2.56	2.56	8.11
MR3	25.00	17.65	100.00	2.56	2.56	2.56	5.26
HR	11.11	14.29	2.56	100.00	17.65	17.65	11.11
CR	8.11	2.56	2.56	17.65	100.00	33.33	11.11
PR	8.11	2.56	2.56	17.65	33.33	100.00	21.21
TW	17.65	8.11	5.26	11.11	11.11	21.21	100.00

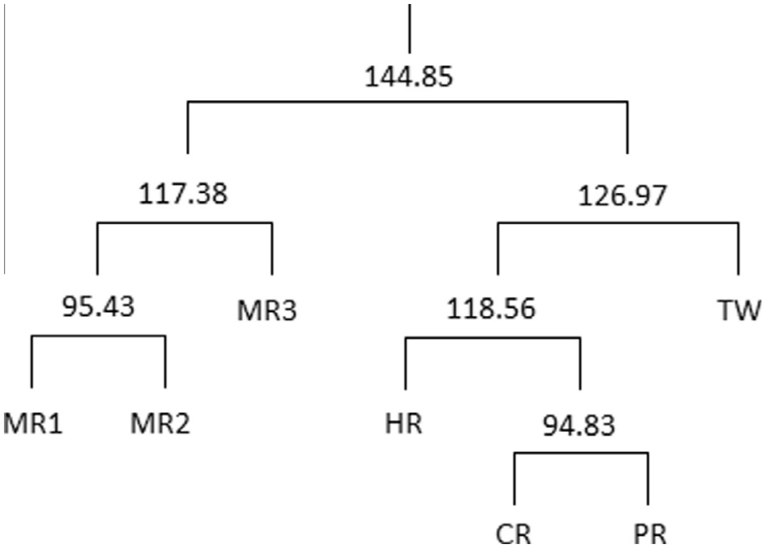


Fig. 4. Clustering of corpora based on idiom usage.

Table 12
Top five most frequently used idioms across the corpora.

Corpus	Idiom 1	Idiom 2	Idiom 3	Idiom 4	Idiom 5
MR1	*fall flat*	*save the day*	*butterflies in stomach*	*jaw drop*	*guilty pleasure*
MR2	*fall flat*	*save the day*	*guilty pleasure*	*devil's advocate*	*jaw drop*
MR3	*guilty pleasure*	*fall flat*	*jaw drop*	*speak volumes*	*close to home*
HR	*chill out*	*mixed feelings*	*go the extra mile*	*lie in*	*last resort*
CR	*come a long way*	*never looked back*	*take it easy*	*happy camper*	*leaps and bounds*
PR	*blockbuster*	*without a hitch*	*happy camper*	*never looked back*	*leaps and bounds*
TW	*chill out*	*lie in*	*take it easy*	*bored to tears*	*hit the sack*

should be scaled down by the percentage of such sentences. Based on our discussion (see Tables 10 and 11), the values of both *p* and *i* are expected to vary across different genres. For example, *p* was found to be over 0.70% and 0.27% on corpora MR3 and TW respectively (note that we only used a list of 580 idioms, so these numbers are expected to be higher). On the other, the classification improvement *i* will vary depending on both the baseline method and the genre. Our sentiment classification experiments were conducted on a genre-neutral corpus of sentences that contain a wide spectrum of uniformly distributed idioms, and as such were used to estimate the expected improvement *i* of 15 to 20 percent points in an unbiased fashion.

7. Conclusions

We have demonstrated the value of idioms as features of sentiment analysis by showing that idiom-based features significantly improve sentiment classification results when such features are present. The overall performance in terms of precision, recall and *F*-measure was improved from 45% to 64% in one experiment, and from 46% to 61% in the other. In terms of *F*-measure, there is an improvement across all classes, but most notably in classifications of positive sentiment. This is mainly attributed to considerably improved recall by 45 percent points in both experiments without compromising the precision. In order to generalize these findings, we combined them with information on idiom distribution. For this purpose, we used corpora commonly used to evaluate performance of sentiment analysis. The improvement of sentiment classification on sentences with idioms should be scaled down by the percentage of such sentences. This number, however, varied across different corpora and requires further research. For example, it was found to be over 0.70% and 0.27% on corpora of movie reviews and tweets respectively.

In addition to experimental results, this study provides resources that can support further research into sentiment analysis including. We created a comprehensive collection of 580 idioms annotated with sentiment polarity, which represents the largest lexico-semantic resource of this kind to utilize in sentiment analysis. In addition, we implemented a set of local grammars that can be used to recognize occurrences of these idioms in text. Thus far, the use of idioms in sentiment analysis was based on *ad hoc* approaches focusing on a severely limited set of idioms that can be identified in text using dictionary lookup methods. Another obstacle in systematically investigating the role of idioms in sentiment analysis is their relative rarity. As explained in previous discussion, corpora commonly used for evaluation of sentiment analysis approaches are biased in their use of idioms, which prevents the findings on the role of idioms in sentiment analysis to be generalized. We assembled a corpus of 2521 sentences with a wide range of idioms used in context. Similarly to idioms themselves, this corpus is also annotated with sentiment polarity, which can be used in systematic evaluation of sentiment analysis approaches that claim to use idioms as features. All resources are freely available for download from the following URL: http://www.cs.cf.ac.uk/idioment.

This is the first in our series of studies on the role of idioms in sentiment analysis. To our best knowledge, this is the first study of this kind in English. Given the positive findings of the initial study, we will be looking to scale up the approach described here. Its main limitation is a significant overhead involved in handcrafting lexico-semantic rules for recognition of idioms and their polarity. To address this bottleneck, we will attempt to automate two crucial steps: (1) encoding local grammars that enable idiom recognition in text, and (2) determining idiom polarity.

The corpus of sentences containing idioms in combination with the set local grammars for idiom recognition can be used as the training set to learn their lexico-syntactic patterns, i.e. induce which idiom components are frozen and which ones vary and in which sense (e.g. inflection, substitution, etc.). These findings can be used to generate lexico-syntactic patterns that will automatically model idioms by means of local grammars. Such rules can be optimized in terms of precision and recall using our corpus annotated with idiom occurrences. Solving a problem of such high complexity would enable utilization of an arbitrary list of idioms without incurring additional implementation overhead.

In order to utilize an arbitrary list of idioms in sentiment analyses, we also need their polarity. In this study, we used crowdsourcing for this purpose. While such an approach proved to be efficient especially when using commercial platforms such as http://www.crowdflower.com/, we would like to fully automate acquisition of idiom polarity as part of scaling up our existing approach. We suggest repurposing existing idiom dictionaries aimed at English language learners. We will conduct sentiment analysis experiments against idiom definitions. The goal of this exercise is to determine whether sentiment analysis results over idiom descriptions (not idioms themselves) are correlated with manual annotations acquired in this study. If this is the case, then it will be possible to generate idiom-based features (i.e. their sentiment polarities) automatically in order to utilize a much wider set of idioms in sentiment analysis.

Combined together, both avenues of further research seek to fully automate the use of idioms in sentiment analysis and minimize the knowledge elicitation bottleneck associated with this task. Finally, taking a long-term view, we have initially annotated both idioms and sentences with a wide range of emotions rather than merely sentiment polarity. This will enable future studies to expand on our existing research to tackle the more complex problem of emotion classification (Spasić et al., 2012).

Acknowledgments

Data cited herein have been extracted from the British National Corpus, distributed by Oxford University Computing Services on behalf of the BNC Consortium. All rights in the texts cited are reserved.

We gratefully acknowledge the help of our colleagues and friends who supported this study by volunteering their time for data annotation. We would like to thank Prof. Mike Thelwall for kindly allowing us to use the Java version of SentiStrength.

References

Attardi, G., & Simi, M. (2006). Blog mining through opinionated words. In *Text Retrieval Conference*. Gaithersburg, Maryland, USA.

Aue, A., & Gamon, M. (2005). Customizing sentiment classifiers to new domains: A case study. In *Proceedings of the international conference on recent advances in natural language processing*. Borovets, Bulgaria.

Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., Goot, E. v. d., Halkia, M., Pouliquen, B., & Belyaeva, J. (2010). Sentiment analysis in the news. In *Proceedings of the international conference on language, resources and evaluation* (pp. 2216–2220). Valletta, Malta.

Beigman Klebanov, B., Burstein, J., & Madnani, N. (2013). Sentiment profiles of multiword expressions in test-taker essays: The case of noun–noun compounds. *ACM Transactions on Speech and Language Processing, 10*, 12.

Bethard, S., Yu, H., Thornton, A., Hatzivassiloglou, V., & Jurafsky, D. (2004). Automatic extraction of opinion propositions and their holders. In *Association for artificial intelligence spring symposium on exploring attitude and affect in text* (p. 2224). Palo Alto, California, USA: The AAAI Press.

BNC Consortium. (2014). The British National Corpus, version 3 (BNC XML Edition), URL: <http://www.natcorp.ox.ac.uk/>.

Breck, E., Choi, Y., & Cardie, C. (2007). Identifying expressions of opinion in context. In *Proceedings of the 20th international joint conference on artificial intelligence* (pp. 2683–2688). Hyderabad, India.

Buczynski, A., & Wawer, A. (2008). Shallow parsing in sentiment analysis of product reviews. In *Proceedings of the LREC workshop on partial parsing: between chunking and deep parsing* (pp. 14–18). Marrakech.

Cohen, W. W. (2014). MinorThird: Methods for identifying names and ontological relations in text using heuristics for inducing regularities from data. URL: <http://www.minorthird.sourceforge.net>.

Damasio, A. R. (1999). *The feeling of what happens: Body and emotion in the making of consciousness*. New York, USA: Harcourt Incorporated.

Darwin, C. (1872). *The expression of emotions in man and animals*. London, UK: John Murray.

Dave, K., Lawrence, S., & Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international world wide web conference*. Budapest, Hungary.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation, 10*, 1895–1923.

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29.

Ekman, P. (1972). Universals and cultural differences in facial expressions of emotions. In J. Cole (Ed.). *Nebraska Symposium on Motivation, 1971* (Vol. 19, pp. 207–283). Lincoln, Nebrasca, USA: University of Nebraska Press.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion, 6*, 169–200.

Esuli, A., & Sebastiani, F. (2006). SentiWordNet: A publicly available lexical resource for opinion mining. In *proceedings of the international conference on language, resources and evaluation* (pp. 417–422). Genoa, Italy.

Everitt, B. S. (1977). *The analysis of contingency tables*. London: Chapman & Hall.

Fraser, B. (1970). Idioms within a transformational grammar. *Foundations of Language, 6*, 22–42.

Ganesan, K., Zhai, C., & Han, J. (2010). Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd international conference on computational linguistics* (pp. 340–348). Beijing, China.

Ganesan, K., Zhai, C., & Viegas, E. (2012). Micropinion generation: An unsupervised approach to generating ultra-concise summaries of opinions. In *Proceedings of the 21st international conference on World Wide Web* (pp. 869–878).

Geertzen, J. (2012). Inter-rater agreement with multiple raters and variable, URL: <https://mlnl.net/jg/software/ira/>, Retrieved May 8, 2014.

Gibbs, R. W., Jr., & Nayak, N. P. (1989). Psycholinguistic studies on the syntactic behavior of idioms. *Cognitive Psychology, 21*, 100–138.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *Technical report, CS224N* (pp. 1–12). Stanford Digital Library.

Grant, L. E. (2005). Frequency of 'core idioms' in the British National Corpus (BNC). *International Journal of Corpus Linguistics, 10*, 429–451.

Greenwood, M., Elwyn, G., Francis, N., Preece, A., & Spasić, I. (2013). Automatic extraction of personal experiences from patients' blogs: A case study in chronic obstructive pulmonary disease. In *Proceedings of the third international conference on social computing and its applications* (pp. 377–382). Karlsruhe, Germany.

Gross, M. (1997). The construction of local grammars. In E. Roche & Y. Schabes (Eds.), (pp. 329–354). The MIT Press.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The Weka data mining software: An update. *ACM SIGKDD Explorations Newsletter, 11*, 10–18.

Jackendoff, R., & Pinker, S. (2005). The nature of the language faculty and its implications for evolution of language (Reply to Fitch, Hauser, and Chomsky). *Cognition, 97*, 211–225.

Krippendorff, K. H. (1980). *Content analysis: An introduction to its methodology*. Beverly Hills, CA: SAGE Publications Inc.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research, 30*, 411–433.

Learn English Today. (2013). URL: <http://www.learn-english-today.com/idioms/idioms_proverbs.html>.

LeDoux, J. E. (1996). *The emotional brain*. New York, USA: Simon & Schuster.

Leech, G. (1993). 100 million words of English. *English Today, 9*, 9–15.

Liu, D. (2003). The most frequently used spoken American English idioms: A corpus analysis and its implications. *TESOL Quarterly, 37*, 671–700.

Liu, B. (2010). Sentiment analysis and subjectivity. In N. Indurkhya & F. J. Damerau (Eds.), *Handbook of Natural Language Processing* (2nd ed.. Boca: CRC Press, Taylor and Francis Group.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 142–150). Portland, Oregon, USA: Association for Computational Linguistics.

Mohammad, S. M., & Turney, P. D. (2010). Emotions evoked by common words and phrases: using mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT workshop on computational approaches to analysis and generation of emotion in text* (pp. 26–34). Los Angeles, California, USA.

Munezero, M. D., Montero, C. S., Sutinen, E., & Pajunen, J. (2014). Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE Transactions on Affective Computing, 5*, 101–111.

Nippold, M. A., & Martin, S. T. (1989). Idiom interpretation in isolation versus context: A developmental study with adolescents. *Journal of Speech, Language and Hearing Research, 32*, 59–66.

Nunberg, G., Sag, I. A., & Wasow, T. (1994). Idioms. *Language, 70*, 491–538.

Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on association for computational linguistics*.

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics* (pp. 115–124).

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*, 1–135.

Rentoumi, V., Vouros, G. A., Karkaletsis, V., & Moser, A. (2012). Investigating metaphorical language in sentiment analysis: A sense-to-sentiment perspective. *ACM Transactions on Speech and Language Processing, 9*, 6.

Rubin, D. C., & Talarico, J. M. (2009). A comparison of dimensional models of emotion: Evidence from emotions, prototypical events, autobiographical memories, and words. *Memory, 17*, 802–808.

Russell, J. (2003). Core affect and the psychological construction of emotion. *Psychological Review, 110*, 145–172.

Shastri, L., Parvathy, A. G. K., Abhishek, Wesley, J., & Blakrishnan, R. (2010). Sentiment extraction: Integrating statistical parsing, semantic analysis, and common sense reasoning. In *Proceedings of the twenty-second conference on innovative applications of artificial intelligence*. Atlanta, Georgia, USA.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., et al. (2013). Recursive deep models for semantic compositionality over a Sentiment Treebank. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 1631–1642). Seattle, Washington, USA: Association for Computational Linguistics.

Spasić, I., Burnap, P., Greenwood, M., & Arribas-Ayllon, M. (2012). A naïve Bayes approach to topic classification in suicide notes. *Biomedical Informatics Insights, 5*, 87–97.

Straessler, J. (1982). *Idioms in English: A pragmatic analysis*. John Benjamins Pub Co.

The Association for the Advancement of Affective Computing. (2014). Emotion annotation and representation language (EARL), Version 0.4.0, 30 June 2006, URL: <http://emotion-research.net/projects/humaine/earl>.

Thelwall, M. (2014). SentiStrength, URL: <http://sentistrength.wlv.ac.uk/>.

Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology, 63*, 163–173.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology, 61*, 2544–2558.

Turney, P., & Littman, M. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems, 21*, 315–346.

Valitutti, A., Strapparava, C., & Stock, O. (2004). Developing affective lexical resources. *Psychology, 2*, 61–83.

Wilson, T., Wiebe, J., & Hoffmann, P. (2009). Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics, 35*, 399–433.

Xie, S.-X., & Wang, T. (2014). Construction of unsupervised sentiment classifier on idioms resources. *Journal of Central South University, 21*, 1376–1384.

Yusifova, G. I. (2013). Syntactic features of English idioms. *International Journal of English Linguistics, 3*, 133–138.