Semantic Topic Compass – Classification Based on Unsupervised Feature Ambiguity Gradation

Amparo Elizabeth Cano(✉), Hassan Saif, Harith Alani, and Enrico Motta

Knowledge Media Institute, The Open University, Milton Keynes, UK
{amparo.cano,h.saif,h.alani,e.motta}@open.ac.uk

Abstract. Characterising social media topics often requires new features to be continuously taken into account, and thus increasing the need for classifier retraining. One challenging aspect is the emergence of ambiguous features, which can affect classification performance. In this paper we investigate the impact of the use of ambiguous features in a topic classification task, and introduce the Semantic Topic Compass (STC) framework, which characterises ambiguity in a topics feature space. STC makes use of topic priors derived from structured knowledge sources to facilitate the semantic feature grading of a topic. Our findings demonstrate the proposed framework offers competitive boosts in performance across all datasets.

Keywords: Topic classification · Feature engineering · Semantics

1 Introduction

Much research focused on understanding what is being discussed on Social Media. From opinion and sentiment mining [16] to event detection [11], one persistent challenge in making sense of this data is the task of assigning topic labels to microposts, which is a necessary step in supervised classification tasks.

Topic characterisation in Social Media poses various challenges due to the event-dependent nature of topics discussed on this outlet. Changes on a topic's representation involve the introduction of event-dependent features, which bring along ambiguous semantic relevance to the topic. For example the word Bataclan, referring to the Bataclan Theatre in Paris is commonly related to Entertainment, however during the November 2015 terrorist attacks in France it became relevant to the Topic Violence. The constant change of a topic's feature space makes apparent the need to be able to characterise the most discriminative features, while identifying ambiguous ones.

Existing feature selection methods such as Information Gain [3] and Odds Ratio [13], assess the problem of feature relevance but perform poorly when a dataset present ambiguous features. More recently, the problem of characterising ambiguous features has been approached using the Ambiguity Measure [12], which enables the selection of the most unambiguous features from a feature set. However such an approach relies on labelled data, and thus renders it less

H. Sack et al. (Eds.): ESWC 2016, LNCS 9678, pp. 350–367, 2016.
DOI: 10.1007/978-3-319-34129-3_22

adequate when modelling topics for social media, where labelling data is costly and becomes rapidly outdated.

In this paper we introduce the Semantic Topic Compass (STC) Framework, which is an unsupervised method that facilitates the semantic feature grading of a topic. This approach relies on the incorporation of feature priors derived from an external corpus to reweigh a Twitter corpus features in an unsupervised manner. Such feature representation partitions a Topic's feature space into four quadrants each representing the level of relevance and ambiguity of a feature to the Topic. To the best of our knowledge none of the existing approaches characterise ambiguity of a topic feature space on unlabelled corpora. The main contributions of this paper can be summarised as follows:

(1) We propose a novel unsupervised approach for topic feature representation based on polar coordinates;
(2) such representation enhances existing ones in characterising features based on both topic relevance and ambiguity;
(3) We propose a weighting strategy that proxies penalties to four feature types characterised by our framework: strongly related, weakly-related, weakly-unrelated and strongly unrelated features.
(4) We evaluate the effectiveness of the proposed framework on a classification task applied over three datasets using both lexical and semantic features.
(5) Our findings demonstrate that the proposed framework offers competitive boosts in performance across all datasets.

2 Related Work

Topic classification on Twitter consist of labelling tweets messages as being either topic-related or topic-unrelated [2]. Most existing works approach this task by training binary machine learning classifiers (e.g., Naive Bayes, SVM) on lexical features extracted either from tweets (i.e., Lexical Features) [10,15,20] and/or from external knowledge sources (i.e., semantic features) [2,6,19]. As such, these works can be divided as lexical approaches and semantic approaches. As for lexical features, Genc et al. [6], proposed the use of unigrams features to map a tweet to the most similar Wikipedia[1] articles which denote the tweets' topic. Sriram et al. [20] classified tweets to a predefined set of topics based on Twitter-specific features such as abbreviations, slangs, user mentions (i.e., @username) and opinionated words.

Rather than relying on lexical features in tweets for topic classification, other approaches proposed enriching the tweets' content with features extracted from external knowledge sources (KS) [2,6,14,19]. For example, [19] mapped a tweet's terms to the most likely resources in the Probbase KS. These resources were used as additional features in a clustering algorithm which outperformed the simple bag of words approach. Muñoz-García et al. [14] proposed an unsupervised vector space model for assigning DBpedia URIs to tweets in Spanish.

[1] http://wikipedia.org.

Cano et al. [2] performed cross-epoch topic classification based on four types of semantic features extracted from DBpedia knowledge graph, including the DBpedia resources, class types, categories and properties of named entities extracted from the tweets.

A persistent issue of both semantic and lexical approaches is the high dimensionality of the feature spaced used for training classifiers, which can reach the order of millions on large Twitter corpora. A large feature space usually affects both, the runtime complexity and the performance of classifiers [9]. To reduce the dimensionality of a feature space, feature selection techniques for topic classification are often used. Feature selection concerns about finding the most discriminative features in a given feature set, aiming at reducing the dimensionality of the classifier's feature space by excluding features of low discrimination power and maintaining high classification performance [5]. Wide range methods have been proposed for automatic selection of features, such as, Information Gain [3], Chi-Squared [5], term frequency and inverse term frequency (TF-IDF) [8], and Odds Ratio [13], etc. Most of these methods function by estimating the probability that a feature belongs to a specific class (topic) and the probability that the feature does not belong to that class. A Common limitation of these methods is that they are not tolerant to imbalanced class distributions in datasets. In other words, they tend to assign high discrimination scores to features that belong to the dominant class in the data (i.e., the class with the highest number of training samples). Also, these methods often perform poorly in the case of ambiguous features, where the presence and absence of a feature with a given class is almost identical. Instead of identifying and filtering out these type of features, they are still assigned a high discrimination score by methods like Odds Ratio, TF-IDF and Chi-Squared.

To address the above limitations Mengle and Goharian [12] proposed the Ambiguity Measure (AM) feature selection method. AM identifies ambiguous features by assigning a higher discrimination score to features pointing to one class than those pointing to more than one class. Their results show that feature selection based on AM outperforms Odds Ratio, Information Gain and Chi-Squared methods. However, similar to these methods AM functions in a supervised fashion, i.e., it requires tweets labelled with their topical orientation. In contrast in this paper we introduce the semantic topic compass framework, which is an unsupervised approach that enables the partition of a topic's feature space characterising ambiguous features. As opposed to previous work, which rely on labelled data for disambiguating features, our proposed approach only relies on topic feature priors extracted from knowledge sources.

3 Ambiguity in Topic Representation

In topic classification, the most discriminative features of a topic are generally those that are semantically-related and semantically-unrelated to the topic. On the other hand, the least discriminative features are those that are weakly-related or weakly-unrelated to the topic, and thus considered ambiguous due to their low discriminative power.

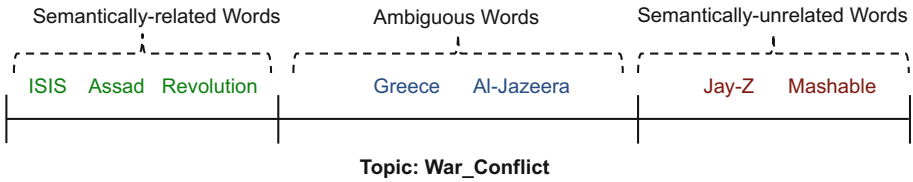

Fig. 1. Spectrum of the semantic level of relatedness/ambiguity of words for the topic War_Conflict

For example, for the topic "`War_Conflict`", depicted in Fig. 1, words such as "`ISIS`"and "`Assad`" are semantically-related to the topic since their underlying semantics (i.e., "`Jihadist_Group`", "`Syria_President`") denote a higher association with the topic "`War_Conflict`". In contrast, the words "`Jay-Z`" and "`Mashable`" are semantically-unrelated to "`War_Conflict`" as their semantics (i.e., "`American_Rapper`" and "`Digital_Media_Website`") are irrelevant to the topic. In between this spectrum lay ambiguous terms, which are not completely relevant nor irrelevant to the topic. For example, the words "`Greece`" and "`Al-Jazeera`" are considered ambiguous as their underlying semantics (i.e., "`Country`", "`News_Agency`") are considered to be weakly associated with the topic "`War_Conflict`" in this example.

Identifying the level of ambiguity of a feature can aid in providing a better representation of a topic. It can also aid in filtering out features keeping only the most discriminatory ones, reducing in this way the dimensionality of the feature space. However to the best of our knowledge there are only few approaches to address the identification of ambiguous features. Moreover existing approaches rely on labelled data (i.e., are supervised) and are only able to discriminate features as being ambiguous or non-ambiguous but they do not differentiate the tendency of the ambiguous word towards the topic (weakly-related or weakly-unrelated).

This paper proposes a novel unsupervised approach to topic feature representation which enables both the relevance/irrelevance feature weighting while giving an ambiguity orientation to a feature. In this paper we propose the use of such topic feature representation to characterise a topic on a topic classification task.

4 Semantic Topic Compass Framework

Since the discriminative power of features used for topic classification relies on the relative use of those features within the topic, we propose to make use of distributional and conceptual semantics to characterise ambiguity in a topic's feature representation. We aim at using such topic characterisation to learn a representation of the topic for classification purposes.

The proposed Semantic Topic Compass (STC) Framework breaks down into three main phases: (1) Semantic Topic Representation: Given a collection of

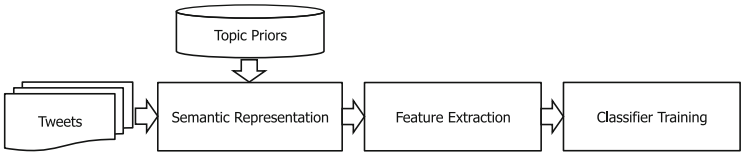

Fig. 2. Pipeline of the proposed topic representation and feature extraction approach.

tweets a semantic representation of a topic is constructed based on the features' semantic relatedness to the topic; (2) Feature Weighting: the semantic representation of the topic is then used to grade and extract features for topic classification.; and (3) Training of a topic classifier: Lastly, the extracted features are used to train the topic classifier.

In the following subsections we describe each of the three steps in the pipeline of our framework in more detail.

4.1 Semantic Representation of Topics' Feature Space

As mentioned before, our semantic topic compass framework relies on incorporating the semantics of words into the feature space of the studied topic, aiming at characterising the relevance and ambiguity of the these features. Hence, this step extracts first the latent semantics of words under a topic, and then incorporates these semantics into the topic's feature space.

Two main approaches have been extensively used in the literature for extracting the semantics of words, namely: the *Distributional Semantic Approach* [4,7] and the *Conceptual Semantic Approach* [18]. The distributional semantic approach (a.k.a statistical semantic approach) relies on the co-occurrence patterns of words in the text for words' semantic extraction, while the conceptual semantic approach makes use of external knowledge sources (e.g., DBpedia) for mapping words with their explicit semantic concepts.

In this paper we investigate the use of both semantic extraction approaches in our topic compass framework. First, describe in this section how to extract and use the distributional semantics of words for space representation and feature grading of a topic. After that, we explain in Sect. 5 how to enrich the space representation of the topic with the conceptual semantics of words.

1. Extracting Words' Distributional Semantics: To extract the distributional semantics of a word, we follow the distributional semantic hypothesis that *words that are used and occur in the same contexts tend to purport similar meanings.*[2] For example, the semantics of the word "`ISIS`" when it occurs with words like "`Kill`" and "`Behead`" denotes that "`ISIS`" refers to the terrorist militia organisation in the Middle East.

Given a tweet collection $\mathcal{T}$ of a topic $\mathcal{P}$ (e.g., "`War_Conflict`"), let's represent each term m in $\mathcal{T}$ (e.g., "`ISIS`") as a vector $\boldsymbol{c} = (c_1, c_2, \ldots, c_n)$ of terms co-occurring with term m in any tweet in $\mathcal{T}$ (e.g., "`Kill`" , "`Behead`" , "`Blood`").

[2] Also known as Statistical Semantics.

We define the degree of correlation between each context word $c_i \in \boldsymbol{c}$ and m based on the *TF-IDF* weighting scheme as follows:

$$corr(m, c_i) = f(c_i, m) \times \log N/N_{c_i} \tag{1}$$

where $f(c_i, m)$ is the number of times c_i occurs with m in tweets, N is the total number of terms, and N_{c_i} is the total number of terms that occur with c_i. Since our main task here is to measure the level of relatedness and ambiguity of a word to the studied topic, we also assign to each context term c_i a topic prior $p(c_i) \in [-1, +1]$, a numerical value representing the initial degree of relatedness of the context term to the topic. Section 4.3 describes how the topic priors of words in the Tweet collection are extracted.

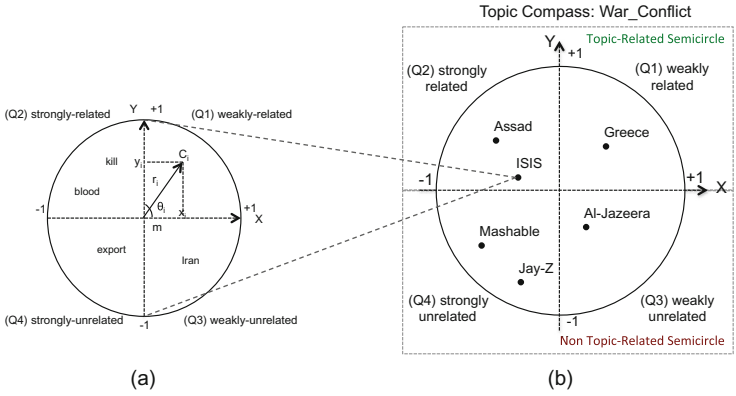

Fig. 3. Semantic representation of: (a) of a term m, and (b) the feature space of the term's related topic $\mathcal{P}$

To extract the collective semantics of the term m we resort to representing the vector $\boldsymbol{c}$ using the polar coordinate system inspired by [16]. In particular, the context vector $\boldsymbol{c}$ is transformed into a 2d circle representation as depicted in Fig. 3, which we will call S_m. The center of this circle represents the target term m and points within the circle denote the context terms of m. The position of c_i is defined jointly by a radius $r_i = corr(m, c_i)$ and an angle measured based on its topic prior $\theta_i = p(c_i)$ as:

$$x_i = r_i \cos\theta_i \qquad\qquad y_i = r_i \sin\theta_i \tag{2}$$

The above representation partitions the context terms of the target term m (e.g., "`ISIS`") into four independent quadrants (Q1, Q2, Q3, Q4) as shown in Fig. 3a. Terms lying on the upper left quadrant (Q2) (e.g., "`kill`" , "`blood`") are strongly related to the topic "`War_Conflict`" , while terms lying on the upper right quadrant (Q1) (e.g., "`Oil`") are weakly-related to the topic. Also, terms residing in the lower left quadrant (Q4) (e.g., "`export`") are strongly-unrelated

to the topic while terms residing on the lower right quadrant (Q3) (e.g., "`Iran`") are weakly-unrelated to the topic.

2. Constructing the Topic's Feature Space: Now we have the semantics of each term m in the tweet collection is represented by a circle S_m. The next step is to derive a global feature space representation for the topic $\mathcal{P}$. To this end, we also use the polar coordinate system, where we represent the topic's feature space as circle S_P, centred at the origin as depicted in Fig. 3b. Each point in the topic's circle S_P denote a term's circle S_m, and is positioned based on the geometric median point of S_m which can be calculated as:

$$g_m = \arg\min_{g_m \in \mathbb{R}^2} \sum_{i=1}^{n} ||c_i - g_m||_2 \tag{3}$$

where the geometric median is a point $g_m = (x_k, y_k)$ in which its Euclidean distances to all the points c_i (context terms) in S_m the is minimum. We can notice that the feature space of the topic (i.e., the topic's Circle S_P) can be also partitioned, in similar way to the circle representation of the terms (i.e., term's circle S_m), into four different quadrants denoting the level of relatedness and ambiguity of the terms under the topic[3].

In the following subsection we show how to use the circle representation of the topic's feature space to weight a tweet for topic classification.

4.2 Feature Weighting for Classifier Training

Representing the feature space of a topic with the proposed framework in the polar coordinate system enhances the standard Euclidean vector space representation in two main aspects: (1) by providing a strength of the relative semantic relevance of a feature to a topic; (2) by augmenting the possible orientations of such relevance to the topic. In this section we propose a method to make use of this information by encoding it into a feature weighting strategy that can be used to weight features in a tweet collection to address a topic classification task.

Let $\mathcal{T}$ be a corpus of tweets denoted as $\mathcal{T} = \{t_1, t_2, .., t_{\mathcal{T}}\}$; where each tweet consists of a sequence of N_t terms denoted by $t = (m_1, m_2, .., m_{Nt})$. Algorithm 1 presents the proposed steps for weighting a tweet t's features based on the topic representation S_P.

The proposed weighting strategy generates a metric that assigns weights to features based on their relevance to the topic[4]. Such relevance is considered based on the position of a feature within the two semicircles (i.e., topic related and non-topic related) described in Fig. 3b. Steps described in Algorithm 1 can be outlined as follows: (i) Given the term frequency vector of a document, iterate over these feature; (ii) For each feature obtain its coordinates in the topic circle

[3] We provide samples of the generated topic circles for the three topics at the following link http://tweenator.com/stc.php.

[4] Features appearing within an axis are considered ambiguous and are smoothed down to a low weight.

Algorithm 1. Feature Weighting based on a Topic's Circle Representation

Input: Term frequency features of t, topic circle S_P, penalties for quadrants (pQ1,pQ2,pQ3,pQ4)
Output: Weighted features for tweet t

for each term $m_i \in t$ **do**
  Extract m_i representation on S_P.
  Compute angle of m_i as $\theta = arctan(y, x)$ in degrees, where x, y are the coordinate representation of m_i in the circle.
  Compute the Euclidean distance of m_i from the circle's origin $(0, 0)$ as $l(m_i) = (x^2 + y^2)^{1/2}$
  if $x > 0 \wedge y > 0$ (first quadrant) **then**
    weight of m_i, $w(m_i) = tf(m_i) * pQ1 * l(m_i)/(180 - angle)/360$
  end if
  if $x < 0 \wedge y > 0$ (second quadrant) **then**
    weight of m_i, $w(m_i) = tf(m_i) * pQ2 * l(m_i)/(angle - 90)/360$
  end if
  if $x > 0 \wedge y < 0$ (third quadrant) **then**
    weight of m_i, $w(m_i) = tf(m_i) * pQ3 * l(m_i)/(angle)/360$
  end if
  if $x < 0 \wedge y < 0$ (fourth quadrant) **then**
    weight of m_i, $w(m_i) = tf(m_i) * pQ4 * l(m_i)/(angle)/360$
  end if
end for

representation; (iii) Based on the coordinates of the feature, weight it considering its magnitude, orientation in the circle and term frequency within the document. The proposed strategy generates a metric which assigns weights from highest to lowest in the following order Q2-Q1-Q3-Q4. Where the highest weight is provided to the strongly-related features (quadrant Q2) and the lowest weight to the strongly unrelated features (quadrant Q4). Both weakly-related and weakly-unrelated features fall close midway within the metric. The proposed penalties for each quadrants enables to emphasize a quadrant's feature or to bring down a quadrant's features relevance. This enables for example to filter out ambiguous features ($pQ1 = 0, pQ2 = 0$), or to highlight the relevance of strongly related features ($pQ2 > 1.0$). This weighting strategy provides a weighted representation of a document that can be used for training a topic classifier.

4.3 Extracting Topic Feature Priors from Semantic Knowledge Sources

In this paper we refer to a topic feature prior as the probability distribution that would express one's beliefs about this feature relevance/irrelevance to a topic before any other evidence is taken into account. Topic feature priors enable us to have a preliminary model of the language related to a topic when no other information about the topic is provided. For example for the topic "`War_Conflict`" such prior information maps violence polarity into violence words such as `looting, war, drugs` and non-violent polarity to background words such as `today, afternoon, happy`.

Word prior lexicon generation relies on the use of a positive and a negative samples of a topic. The feature prior representation of a Topic consists on getting all features (e.g., words) of a topic dataset and assigning to each of them a weight representing how well the feature is relevant to the topic. Social knowledge sources provide a rich textual information covering a large number of topics. In this work we use as a positive sample of a topic the set of articles' abstracts

belonging to categories and subcategories derived for a topic in DBpedia. As a negative sample we use a set of tweets which are not related to this topic[5]. So feature priors for the topic war for example would look like (feature:explosion war: 0.8 non-war: 0.2; feature: sandwich war: 0.02 non-war: 0.98; and so on for each feature in the War corpus).

To derive lexical features we use bag of words over the dataset. To derive semantic features we extract and disambiguate entities appearing on these abstracts, using AlchemyAPI. We then SPARQL queried DBpedia to obtain specific semantic features about each entity e.g., categories, class type. Based on these datasets to derive topic feature priors we employ the widely-used information gain method to select highly discriminative words under each class.

5 Conceptual Semantic Enrichment

In the previous sections we showed our proposed framework to facilitate feature grading of topics using the words' distributional semantics. However, using the distributional patterns of a word (i.e., word's context) to detect its semantics in tweets is sometimes insufficient. For example, the word "`ISIS`" in "`ISIS continues spreading like a malignant tumor!`" lack enough context to determine its semantics. Nonetheless, existing knowledge sources provide a wealth of structure data that can be used to address this issue. For example the word "`ISIS`" is a resource in DBpedia associated with the semantic category "`Jihadist_Group`". Such association denotes a stronger relatedness with the topic "`War_Conflict`". To account for semantic relatedness we propose to enrich our topic compass framework, with the explicit or conceptual semantics of words in tweets. To this end, we follow two main steps:

1. Entity Extraction and Semantic Mapping: This step extracts named entities appearing in a tweet collection (e.g., "`ISIS`", "`Bashar_Al_Assad`", "`Barack_Obama`") using the semantic extraction tool, AlchemyAPI.[6] Then, each entity (e.g., "`Bashar_Al_Assad`") is mapped to a (i) semantic concept provided by AlchemyAPI (`Alc:Person`); (ii) DBpedia Category (`dbc:Presidents_of_Syria`); and (iii) DBpedia Class (`dbo:Arab_--Politician`), shown in Table 1.

2. Conceptual Semantic Enrichment: This step incorporates the conceptual semantics extracted from the previous step into the semantic representation of the topics' feature space. As mentioned in Sect. 4.1, the context of a term m is represented as a vector $\boldsymbol{c} = (c_1, c_2, \ldots, c_n)$ of terms that occur with m in a given tweet collection. Our semantic enrichment is done on this vector as follows:

– For AlchemyAPI Concepts, we extend the contextual vector $\boldsymbol{c}$ with the semantics $\boldsymbol{s} = (s_1, s_2, \ldots, s_m)$ of named entities $\boldsymbol{e} = (e_1, e_2, \ldots, e_m)$ that occur with m in the tweet collection as:

$$\boldsymbol{c_s} = \boldsymbol{c} + \boldsymbol{s} = (c_1, c_2, \ldots, c_n, s_1, s_2, \ldots, s_m) \tag{4}$$

[5] Notice that the tweet sample used for deriving priors is independent of the corpus used for topic classification in the experiments section.

[6] www.alchemyapi.com.

Table 1. Example of named entities extracted from tweets and mapped to their associated AlchemyAPI Concept, DBpedia Category, and DBpedia Class

Entity	Alchemy concept	Dbpedia category	Dbpedia class
ISIS	Organization	Jihadist_Groups	Populated_Place
Barack_Obama	Person	Presidents_of_the_US	Politician
Syria	Country	Middle_Eastern_Countries	Location
Bashar_Al_Assad	Person	Presidents_of_Syria	Arab_Politician

– For DBpedia Categories and Classes, we replace the entire contextual vector $\boldsymbol{c}$ with the semantic categories $\boldsymbol{o} = (o_1, o_2, \ldots, o_m)$ or the semantic classes $\boldsymbol{l} = (l_1, l_2, \ldots, l_m)$ of the entities in $\boldsymbol{e}$ as:

$$\boldsymbol{c_o} = \boldsymbol{o} = (o_1, o_2, \ldots, o_m) \tag{5}$$

$$\boldsymbol{c_l} = \boldsymbol{l} = (l_1, l_2, \ldots, l_m) \tag{6}$$

where $\boldsymbol{c_s}$, $\boldsymbol{c_o}$ and $\boldsymbol{c_l}$ are the new semantically-enriched contextual vectors of m, which will be subsequently used instead of $\boldsymbol{c}$ to extract the semantic circle representation of m as described in Sect. 4.1. It is worth noting that the semantic enrichment done through Eqs. 5 and 6 results in topics' feature spaces completely represented by the entities' semantic categories or classes. Conversely, the feature spaces inferred from Eq. 4 are mix of words, named entities and AlchemyAPI concepts. The reason behind this representation variation is twofold. First, investigate the impact of words' semantics when solely used in our framework for feature grading. Secondly, unlike the large variety of DBpedia categories and types, the number of distinct concepts retrieved by AlchemyAPI from our datasets is limited to 41 concepts only. Relying on these concepts in our framework leads to sparse feature space representation of topics, which often results in low topic classification performance [17].

6 Experimental Setup

Here we present the experimental set up used to assess our proposed topic compass framework. We evaluate the effectiveness of our STC framework in a topic classification task. Specifically, we apply our framework on different Twitter datasets for feature extraction and grading. Then, the extracted features are used to train supervised classifiers for topic classification. Thus, our evaluation setup requires the selection of (i) Twitter datasets for feature extraction, (ii) baselines methods for cross-comparison, and (iii) the knowledge source from which the topic's prior are extracted. All these elements will be explained in the following subsections.

6.1 Datasets

To assess the performance of the classification task we require the use of datasets annotated with a topic label. For this work we selected three evaluation datasets, previously used in the literature of topic classification on Twitter [2]. These datasets consist of a collection tweets of Violence-related topics: *Disaster_Accident*, *Law_Crime* and *War_Conflict*. Tweets in each dataset are manually labelled with negative and positive scores denoting their relatedness to the topic.[7] Size and number of word unigrams, within each dataset are summarised in Table 2.

Table 2. Statistics of the three datatsets used for evaluation

Dataset	Tweets	Unigrams	Categories	Classes
Disaster_Accident	2,528	6,341	4,522	124
Law_Crime	1,967	4,540	3,582	113
War_Conflict	1,939	4,502	3,533	110

6.2 Baselines

As mentioned in the Sect. 2 different types of lexical and semantic features have been used in multiple works on topic classification. In this paper we choose to compare the features extracted by the STC framework against the following state-of-the-art lexical and semantic feature types.

Lexical Feature Baselines

TF Features: denoting word unigrams weighted by their term frequency in the tweets.
TF-IDF Features: denoting word unigrams weighted by using term frequency inverse document frequency.
LDA Features: referring to word unigrams weighted by the latent topic extracted from tweets using the probabilistic generative model, LDA [1]. To extract these latent topics from our datasets we use an implementation of LDA provided by Mallet.[8] LDA requires defining the number of topics to extract before applying it on the data. We experimented with different numbers of topics. Among all choices, 10 topics was the optimal number giving the highest classification performance for this baseline.
AM Features: referring to features weighted based on the Ambiguity Measure feature selection method [12].

[7] Details about the construction and the annotation of these datasets are provided in [2].
[8] http://mallet.cs.umass.edu/.

Semantic Feature Baselines

DBpedia Features: refer to two different types of semantic features obtained from DBpedia: (i) Semantic Categories (Cat) and (ii) Semantic Classes (Cls). To extract these features, we first extract the named entities in the Twitter datasets and map them after that to their classes and categories in DBpedia. Table 2 shows the number of semantic categories and classes extracted from each dataset.

AlchemyAPI Concepts: this type of features refers to the semantic concepts appearing in tweets. We extract these features using the Alchemy semantic extraction service. The number of the unique concepts extracted from our datasets is 41.

Examples of the above three types of semantic features are provided in Table 1.

7 Evaluation and Results

In this section we report the evaluation results obtained from using the features extracted by our STC framework in topic classification task. To this end, we use Naive Bayes classifiers. Our baselines of comparison are classifiers trained from the 7 types of lexical and semantic features described in Sect. 6.2. Results in all experiments are computed using 2-fold cross validation over 5 runs of different random splits of the data to test their significance. Statistical significance is done using the T-Test.

Evaluation in the subsequent sections consists of 4 main steps:

1. Investigate the impact of the feature weighting in our STC framework on the classification performance (Sect. 7.1).
2. Measure and compare performance of the STC framework against other supervised and unsupervised feature representation and selection models (Sect. 7.2).
3. Study the effect of enriching the STC framework with conceptual semantics on the topic classification performance (Sect. 7.3).

7.1 Feature Weighting with the STC Framework

The first task in our evaluation is to assess the performance of our semantic feature grading framework. As described in Sect. 4, STC provides an unsupervised approach for weighting a topic feature space. In this section we investigate how such topic representation along with the weighting strategy presented in Algorithm 1 performs in a classification task. Table 3 shows the results of binary topic classification performance for the three datasets following the weighting approach of the proposed STC framework. The table reports four sets of precision (P), recall (R), and F1-measure (F1), one for each dataset, and the fourth one shows the averages of the three.

The first column of the Table presents the penalties assigned to the four quadrants in Algorithm 1. When these penalties are higher or lower than 1.0 they enable to highlight or lessen respectively the weights assigned to features on a particular quadrant. We first analysed a base setting in which all penalties are set to 1.0. In this setting all weights derived directly from the topic circle are kept except for those situated on the axis which are smoothed down. This setting yields a consistent significant boost in P-measure on the three datasets with an increment in P of 7.36 % over the TF baseline (t-test with $\alpha < 0.01$). High precision in this setting shows the effectiveness of the topic circle approach to distribute topic independent features within the axis, which aids in improving the topic classification task.

Table 3. Performance of the TF and STC based classifiers. Tuples on the left side of the second section of the table represent the weights assigned to penalties $pQ1, pQ2, pQ3, pQ4$ respectively. The values highlighted in bold correspond to the best results obtained for each topic. A $\star$ denotes that the F-measure of a given weighted feature significantly outperforms the corresponding TF baseline. Significance levels: *p-value* < 0.01.

	Diss_Acc			Law_Crime			War_Conflict			Average		
	P	R	$F1$	P	R	$F1$	P	R	$F1$	P	R	$F1$
TF	0.8634	0.8886	0.8758	0.8245	0.882	0.8523	0.8205	0.8892	0.8535	0.8361	0.8866	0.8605
1.0,1.0,1.0,1.0	0.9383	0.7613	0.8405	0.8966	0.8380	0.8661	0.8889	0.8259	0.8558	0.9079$\star$	0.8084	0.8541
0.0,1.0,0.0,1.0	0.9024	0.6783	0.7744	0.8413	0.7573	0.7965	0.8141	0.6387	0.7156	0.8526	0.6914	0.7621
0.0,1.0,1.0,0.0	0.8826	0.8843	0.8835	0.8468	0.8699	0.8581	0.8466	0.8865	0.8660	0.8586	0.8802	0.8692$\star$
0.1,1.0,0.1,1.0	0.9024	0.7130	0.7965	0.8624	0.7636	0.8098	0.8276	0.7005	0.7585	0.8641	0.7257	0.7882
2.0,1.0,2.0,1.0	0.9358	0.8075	0.8669	0.8620	0.8689	0.8651	0.8457	0.8728	0.8588	0.8811	0.8497	0.8636$\star$
0.1,2.0,0.1,2.0	0.9043	0.7048	0.7921	0.8530	0.7527	0.7996	0.8183	0.6832	0.7446	0.8585	0.7135	0.7787
2.0,2.0,1.0,1.0	0.9400	0.7879	0.8572	0.8353	0.8819	0.8579	0.8372	0.8723	0.8542	0.8708	0.8473	0.8564
2.0,2.0,2.0,2.0	0.9317	0.8118	0.8676	0.8679	0.8711	0.8693	0.8709	0.8696	0.8698	0.8901$\star$	0.8508	0.8689$\star$

The last section of Table 3 presents results for different penalty settings. In particular we find that keeping weights on quadrant 2 (strongly-related) and 4 (strongly unrelated) while lessening Q1, Q3 (ambiguous quadrants) (i.e., setting `(0.0,1.0,0.0,1.0)`) boost performance but decreases recall. This result indicates the importance of ambiguous terms for a classifier to learn how to discriminate relevant from irrelevant topics.

We also find that highlighting the irrelevance of weakly-unrelated features (Q3) while keeping the strongly-related (Q2) weights and smoothing those features which are strongly unrelated (Q4) and weakly-related (Q1) (i.e., `(0.0,1.0,1.0,0.0)`) provides the best boost in F measure when compared against the TF baseline. This setting improves P in 2.25 % in average over the TF baseline (t-test with $\alpha < 0.01$). This result stresses the importance of balancing weights between the two types of ambiguous features identified by this framework. In particular we find that stressing the irrelevance of the weakly-unrelated features by smoothing down the weakly-related aids in improving performance. Given that this latter setting provides the best boost in F-measure we keep

this setting to perform a cross comparison of the different type of features and weighting baselines in the following section.

7.2 Cross Comparison Results

Here we evaluate the performance of STC against both, the lexical and semantic baselines described in Sect. 6.2. The first section of Table 4 shows the performance obtained with the lexical baselines: (1) Term frequency (TF); (2) TF-IDF; (3) LDA; and (4) Ambiguity metric (AM). All TF-IDF, LDA and AM offer competitive results to the standard TF metric for all datasets. Specifically, TF-IDF offers an overall boost in P; however it's the LDA baseline the one that consistently outperforms the TF baseline on both P and F1 in all datasets. The second section of Table 4 shows the performance of the semantic baselines: (5) Semantic Categories (Cat), (6) Semantic Classes (Cls), and (7) AlchemyAPI Concepts (Alc). In particular, the Cat baseline consistently outperforms in P all lexical baselines in the datasets.

Average results for the lexical baselines (Avg_{Lex}) and the semantic baselines (Avg_{Sem}) in Table 4 show that the semantic baselines slightly outperform the lexical ones, but give lower performance in R and F1. Unlike feature weighting in the lexical baselines which considers all terms in the datasets, the feature weighting in the semantic baselines considers the named-entities only (see Sect. 6.2). This might explain the low recall and F1 of the semantic baselines.

From the set of baselines, Cat offers the best performance, while LDA offers the best boost in F1 across datasets. The third and fourth sections of Table 4 present results for the STC framework with distributional and conceptual semantics respectively. For the first case we present two weighting settings: (i) The default setting STC_{Def}, where all penalties are set to 1.0 (1.0, 1.0, 1.0, 1.0), and the STC_{Bal} setting (0.0, 1.0, 1.0, 0.0) which has shown to yield a balanced performance in P, R, and F1 among other settings, as described in Sect. 7.1.

Here STC_{Def} consistently outperform all 7 baselines in P across all datasets, with an average boost of 7 % (significant at $p < 0.01$) when compared to TF and 4 % (significant at $p < 0.01$) when compared to the highest baseline in P (Cat). In particular STC_{Def} provides best results for Law_Crime with a boost in P of 7.4 % when compared with its TF baseline and of 5.8 % when compared to its highest lexical baseline (LDA). STC_{Def} setting, leaves out features lying on the axes, which can also be considered as ambiguous features, this might explain the boost in precision. STC_{Bal} offers competitive results for P outperforming all lexical baselines, however the semantic baseline Cat outperforms it. Nonetheless, STC_{Bal} shows a consistent but slight boost in F1 for all datasets. The STC_{Bal} removes strongly unrelated (Q4) and weakly-related (Q1) features, this might explain the boost in P and F, however it also shows that removing these features can impact performance in R.

We believe that the above results show the effectiveness of our STC framework for feature grading over the baselines reported in this paper. While STC_{Def} and STC_{Bal} consistently boost P and F1 respectively across all dataset, each

Table 4. Cross-comparison results of the STC framework against the lexical and semantic baselines. The values highlighted in bold correspond to the best results obtained for each topic. A $\star$ denotes that the F-measure of a given weighted feature significantly outperforms the baselines. Significance levels: p-value < 0.01. The values highlighted in bold correspond to the best results obtained for each case.

	Diss_Acc			Law_Crime			War_Conflict			Average		
	P	R	$F1$	P	R	$F1$	P	R	$F1$	P	R	$F1$
TF	0.8634	0.8886	0.8758	0.8245	0.8820	0.8523	0.8205	0.8892	0.8535	0.8361	0.8866	0.8605
TFIDF	0.8702	0.8693	0.8697	0.8449	0.8467	0.8457	0.8397	0.8588	0.849	0.8449	0.8467	0.8457
LDA	0.8797	0.8876	0.8836	0.8368	0.8790	0.8573	0.8380	0.8910	0.8637	0.8515	0.8858	0.8682
AM	0.8587	0.8891	0.8736	0.816	0.8726	0.8433	0.8145	0.8866	0.849	0.8297	0.8827	0.8553
Avg_{Lex}	0.868	0.8836	0.8756	0.8305	0.8700	0.8496	0.8281	0.8814	0.8538	0.8422	0.8783	0.8597
Cat	0.9015	0.6238	0.7372	0.8500	0.5656	0.6791	0.8519	0.6692	0.7493	0.8678	0.6195	0.7219
Cls	0.8627	0.8211	0.8414	0.8164	0.8281	0.8221	0.8091	0.8324	0.8206	0.8294	0.8272	0.8280
Alc	0.8763	0.8789	0.8776	0.8372	0.8730	0.8547	0.8408	0.8797	0.8598	0.8514	0.8772	0.8640
Avg_{Sem}	0.8802	0.7746	0.8187	0.8345	0.7556	0.7853	0.8339	0.7938	0.8099	0.8486	0.7746	0.8046
STC with Distributional Semantics												
1.0,1.0,1.0,1.0	0.9383	0.7613	0.8405	0.8966	0.8380	0.8661	0.8889	0.8259	0.8558	0.9079$\star$	0.8084	0.8541
,0.0,1.0,1.0,0.0	0.8826	0.8843	0.8835	0.8468	0.8699	0.8581	0.8466	0.8865	0.8660	0.8586	0.8802	0.8692$\star$
STC with Conceptual Semantics												
STC_Cat	0.8456	0.8986	0.8713	0.7815	0.9145	0.8427	0.7899	0.8827	0.8336	0.8056	0.8986$\star$	0.8492
STC_Cls	0.86	0.8672	0.8635	0.7967	0.8669	0.8302	0.79	0.8817	0.8332	0.8152	0.8720	0.8422
STC_Alc	0.9244	0.7814	0.8469	0.8299	0.8866	0.8571	0.8202	0.8741	0.8459	0.8581$\star$	0.8473	0.8499

dataset has a different boost for each setting, this is expected since ambiguity and specificity are topic-dependent features. The following section presents results for STC with conceptual semantics.

7.3 Evaluation of the STC Framework with Semantic Enrichment

Here we evaluate the impact of the semantic enrichment when applying the STC framework. The last section of Table 4 presents results for STC with conceptual semantics for all datasets. In average STC_Cat boosts R with 19.1 % over the Cat baseline offering on average a slight boost of 1.2 % over the highest R baseline (TF). Considering the independent results in Table 4 we see that STC_Cat offers a boost in R for all datasets. In particular it provides the highest boost for Law_Crime with 3.35 % over TF, which provides the highest baseline.

The STC_Cls and STC_Alc also offer competitive baselines for P however they don't outperform the best results obtained with STC with distributional semantics. While the semantic enrichment improves upon the baselines in recall, it is the STC with distributional semantics the feature that provides the overall best performance in F1, outperforming also the baselines in P and offering a competitive R.

8 Discussion

In this paper we introduced a novel approach for topic feature representation and weighting which enhances the state-of-the-art feature weighting approaches in providing an ambiguity orientation to each feature. This approach facilitates the identification and filtering of relevant, weakly-related, weakly-unrelated, and unrelated features of a topic's feature space. The geometric nature of the proposed approach facilitates the partition of the feature space, allocating ambiguous feature over the axes and over quadrants 1 and 3.

In order to discuss the effect of ambiguous features in the classification performance task we performed a correlation analysis over gain in performance. For this analysis we focus on the gain provided by the (0.0, 1.0, 0.0, 1.0) setting which lessens the relevance of ambiguous features while highlighting the strongly relevant and strongly irrelevant features. We computed Pearson's correlation between the gain in P, R and F-measure of this setting for lexical and semantic feature versus their corresponding following ratios: (1) ratio of number of weakly-related (WR) to strongly-related (SR) features (WR/SR); (2) ratio of number of weakly-unrelated (WU) to strongly-related (SR) features (WU/SR); (3) ratio of number of the sum of WR and WU to SR ($WR + WU/SR$).

The computed correlations for these ratios are presented in Fig. 4 (statistically significant at $p < 0.05$). These results show the impact of filtering/keeping weakly-related or weakly-unrelated features in boosting performance on a classification task. In particular they reveal the compromise of the use of ambiguous features on this task. Lowering the weight of weakly related features has a slightly positive impact on Precision, while having a positive moderate effect in increasing Recall. Moreover the correlation analysis show that lessening the effect of both WR and WU has a high positive effect in increasing F-measure on the classification task (significant at $p < 0.05$).

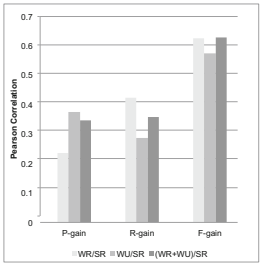

Fig. 4. Pearson correlation for P, R, F-measure gain versus the following ratios: weakly-related (WR) to strongly-related (SR) (WR/SR); weakly-unrelated (WU) to strongly-related (SR) (WU/SR) and ($WR + WU/SR$). Correlation windows: Negligible ($0 - 0.19$); Weak ($0.2 - 0.39$); Moderate ($0.4 - 0.69$); High (>0.69).Statistically significant at $p < 0.05$.

In the previous section we demonstrated the positive effect of the use of STC framework in improving performance upon the TF weighting scheme. We also show that the use of distributional semantics improve performance over the lexical baselines. While Category features weighted with the STC improved upon both STC with distributional semantics and baselines in Recall. This is an expected results since the use of semantic categories provides a generalisation over the type of entities contained on a tweet. However, in our results semantic features did not outperformed in P and F-measure. When computing the density of quadrants for these features we observed that for both Cat and

Cls, the percentage of features appearing on Q1 and Q3 is less than 3 %. In this case lessening the ambiguity of those features does not have an apparent effect on the classification performance.

9 Conclusions

In this paper we introduced the Semantic Topic Compass Framework (STC) which enables to characterise the orientation of features towards the relevancy/irrelevancy of a topic. STC is an unsupervised approach relying only on the use of topic feature priors derived from semantic knowledge sources. It is based on the use of distributional and conceptual semantics to characterise feature ambiguity using a polar representation of a topic's feature space. Based on such feature representation we proposed a weighting strategy which encodes both ambiguity orientation and topic relevance. The proposed strategy proved useful in the topic classification task. To the best of our knowledge this is the first approach to address feature ambiguity characterising a feature's ambiguity-orientation towards being relevant/irrelevant to a topic. In particular our results show that there is a compromise between the use and filtering of weakly related and weakly unrelated features. Future work includes to iterate the process of characterising ambiguity within an active learning setting.

Acknowledgment. This work was supported by the EU-FP7 project SENSE4US (grant no. 611242) and the UK HEFCE project MK:Smart.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
2. Cano, A.E., He, Y., Alani, H.: Stretching the life of twitter classifiers with time-stamped semantic graphs. In: Mika, P., et al. (eds.) ISWC 2014, Part II. LNCS, vol. 8797, pp. 341–357. Springer, Heidelberg (2014)
3. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley, New York (2012)
4. Firth, J.R.: A Synopsis of Linguistic Theory. Studies in Linguistic Analysis (1930–1955) (1957)
5. Galavotti, L., Sebastiani, F., Simi, M.: Experiments on the use of feature selection and negative evidence in automated text categorization. In: Borbinha, J.L., Baker, T. (eds.) ECDL 2000. LNCS, vol. 1923, pp. 59–68. Springer, Heidelberg (2000)
6. Genc, Y., Sakamoto, Y., Nickerson, J.V.: Discovering context: classifying tweets through a semantic transform based on wikipedia. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) FAC 2011. LNCS, vol. 6780, pp. 484–492. Springer, Heidelberg (2011)
7. Harris, Z.S.: Distributional structure. Word **10**(2–3), 146–162 (1954)
8. How, B.C., Narayanan, K.: An empirical study of feature selection for text categorization based on term weightage. In: Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 599–602. IEEE Computer Society (2004)

9. Janecek, A., Gansterer, W.N., Demel, M., Ecker, G.: On the relationship between feature selection and classification accuracy. J. Mach. Learn. Res. **4**, 90–105 (2008)
10. Lee, K., Palsetia, D., Narayanan, R., Patwary, M.M.A., Agrawal, A., Choudhary, A.: Twitter trending topic classification. In: 2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW), pp. 251–258. IEEE (2011)
11. McCreadie, R., Macdonald, C., Ounis, I., Osborne, M., Petrovic, S.: Scalable distributed event detection for twitter. In: Proceedings of the 2013 IEEE International Conference on Big Data, Santa Clara, pp. 543–549, 6–9 October 2013
12. Mengle, S.S., Goharian, N.: Ambiguity measure feature-selection algorithm. J. Am. Soc. Inf. Sci. Technol. **60**(5), 1037–1050 (2009)
13. Mladeni'c, D., Grobelnik, M.: Feature selection for classification based on text hierarchy. In: Text and the Web, Conference on Automated Learning and Discovery CONALD-98. Citeseer (1998)
14. Muñoz-García, O., García-Silva, A., Corcho, O., Higuera Hernández, M., Navarro, C.: Identifying topics in social media posts using dbpedia (2011)
15. Phan, X.H., Nguyen, L.M., Horiguchi, S.: Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: Proceedings of the 17th International Conference on World Wide Web, pp. 91–100. ACM (2008)
16. Saif, H., Fernandez, M., He, Y., Alani, H.: SentiCircles for contextual and conceptual semantic sentiment analysis of twitter. In: Presutti, V., d'Amato, C., Gandon, F., d'Aquin, M., Staab, S., Tordai, A. (eds.) ESWC 2014. LNCS, vol. 8465, pp. 83–98. Springer, Heidelberg (2014)
17. Saif, H., He, Y., Fernandez, M., Alani, H.: Semantic patterns for sentiment analysis of twitter. In: Mika, P., et al. (eds.) ISWC 2014, Part II. LNCS, vol. 8797, pp. 324–340. Springer, Heidelberg (2014)
18. Sheth, A., Ramakrishnan, C., Thomas, C.: Semantics for the Semantic Web. Idea Group Publishing, p. 1 (2005)
19. Song, Y., Wang, H., Wang, Z., Li, H., Chen, W.: Short text conceptualization using a probabilistic knowledgebase. IJCAI **3**, 2330–2336 (2011)
20. Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., Demirbas, M.: Short text classification in twitter to improve information filtering. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 841–842. ACM (2010)