ABSTRACT

Citation in a research paper is a reference to a published source. The citing paper is often referred to as the source paper, and the cited one is referred to as the target paper. Citation helps us to identify the important links to the work done in the paper. Sentimental analysis of citations in scientific papers is a new interesting problem due to many differences in genres. In this project, we focus on automatic identification of positive and negative sentiment polarity in citations of a research paper. This helps a researcher to identify the useful works in the research paper based on positive citations. Negative citations help the researcher to identify the gaps in a research field.

Key words: Machine Learning, Sentimental Analysis, Citations, Support Vector Machine, Naïve Bayes

CONTENTS

LIST OF FIGURES

CHAPTER-1: INTRODUCTION

The growing availability of textual data on the World Wide Web has triggered an increase in research activities focusing on this area. As a result, research on opinion-centric information retrieval continues to be of interest for many applications, ranging from determining product ratings based on consumer reviews for finding out the political stance of individuals and communities.

A popular example of such an application is trying to predict the sentiment of a movie review. With development of sites such as the Internet Movie Database (IMDB1), it is possible to gather information about the sentiment of reviewers on the website towards a movie, from the rating assigned by them. The website allows each reviewer to rate a movie on the scale of 1 to 10 (10 being highest), which can be viewed by the later visitors to the website in an aggregated form. While the exact algorithm to calculate this aggregate score is not disclosed in order to avoid any malicious attempts to temper the rating score, the website administrators have acknowledged that the aggregation is performed as a weighted vote average2. It is thus trivial to calculate this aggregate score given all ratings by individual reviewers once we have the appropriate weights as it is a function of those ratings. However, in the absence of any ratings provided by individual reviewers, calculating the aggregate score becomes a problem which is difficult to tackle. This problem is formally known as sentiment analysis.

Sentiment analysis (also known as opinion mining or emotion AI) refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information.

This project aims to identify the positive and negative sentiment polarity of citations to research papers. Though this sounds quite similar to the sentence- based sentiment analysis, the sentence-based sentiment analysis does not work quite well in this case due to numerous reasons.

To list a few of them-

a) Sentiments are not explicitly depicted in most citations because of two major reasons. Firstly, citations mostly describe an approach or state a fact. Secondly, works are cited mostly in a neural way.

b) The lexical items used to express sentiments is different in scientific texts and general sentences.

c) The scope of influence of citations varies widely from a single clause to several paragraphs.

Tracking citations is an important component of analysing scholarly big data. Citations provide a quantitative way to measure the quality of published works, to detect emerging research topics, and to follow evolving ones.

In this work, we argue that not all citations are equal. While some indeed indicate that the cited work is used or, more importantly, extended in the new publication, some are less important, e.g., they discuss the cited work in the context of related work that does not directly impact the new effort.

Citation sentiment can be used for determining the quality of a paper for ranking in citation indexes by including negative citations in the weighting scheme. Ranking algorithms, such as PageRank, have already been extended to include negative edge weights for different domains. For bibliometrics, however, the effects of these negative edges have yet to be analysed. This information can be used to not only rank articles on a global level for determining their importance, but also provide personalized ranking and recommendations of articles which have been tailored to a particular researcher.

Typically, it takes several years within a field, going to conferences and reading publication over publication to get an intuition about the people involved and their connections and contributions. Luckily, with some python and its extensive libraries, we can speed up this process and quickly generate insight into this network by simply analysing all relevant journal publications. This can be done using network analysis.

Granularity becomes particularly important in the case of product reviews, where different features of a product have to be examined. In such cases, sentiment is determined with respect to the target objects. Product features can be thought of as individual target objects towards which the sentiment is to be detected. For example, a camera with an overall positive review may have low battery life. This task has been tackled by many researchers with varying degrees of success (Hu and Liu, 2004; Liu et al., 2005; Wu et al., 2009; Dasgupta and Ng, 2009). This can be mapped to my research problem, where a single sentence can contain citations to many different papers. In such cases, the sentiment toward only the target paper is to be determined while the rest of the sentiment, however strongly expressed, needs to be ignored.

1.2 Supervised Learning Methods

Supervised learning as the name indicates the presence of a supervisor as a teacher. Basically, supervised learning is a learning in which we teach or train the machine using data which is well labelled that means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples(data) so that supervised learning algorithm analyses the training data (set of training examples) and produces a correct outcome from labelled data.

The increasing availability of labelled data has played an important role in the application of supervised machine learning methods to sentiment analysis. These methods represent the labelled data in the form of a set of features. The features are then used to learn a function for classification of unseen data. In this dissertation, I approach the problem of sentiment analysis as a classification task.

Supervised learning can also be performed using multiple classifiers, particularly if the labelling scheme allows for hierarchical relations. As described earlier, one example of this is the work by Pang and Lee (2004). They represented sentences in the given document as graph nodes and calculated the minimal cut on that graph to identify the subjective sentences. Afterwards, standard machine learning classification algorithms (NB and SVMs) were applied only on the extracted subjective sentences to predict their polarity. On a balanced polarity corpus of 2,000 reviews, the minimum-cut framework resulted in an accuracy of 86.4%, which represents a statistically significant improvement in polarity classification accuracy over previous attempts.

Another example of supervised sentiment classification at a finer granularity is the work by Wilson et al. (2009), who proposed a system for automatically distinguishing between prior and contextual polarity at the sentence level. The key idea is that the polarity of words can change on the basis of other words present in their context. For instance, in the phrase National Environment Trust, the word trust has positive priors but is used here in a neutral context. Starting with a collection of clues marked with prior polarity, the so-called contextual polarity in the corpus was identified as positive, negative, neutral or both. A two-step approach was taken by first classifying each phrase containing a clue as either neutral or polar. In the second step, only the phrases marked as polar were considered further. A corpus of 8,984 sentences was constructed from 425 documents. Unlike movie reviews, these documents did not contain any information for the automatic extraction of labels. Therefore, this corpus was annotated manually by examining each sentence. As a result, 15,991 subjective expressions were found in the 8,984 sentences. 66 documents (1,373 sentences/2,808 expressions) were separated as the development test and the remaining data was used as the test set. For the task of neutral-polar classification, 28 different features based on word context, dependency parse tree, structural relationship, sentence subjectivity and document topic we used in a machine learning framework. For the task of polarity classification, ten features based on word tokens, word prior polarity, as well as presence of negation, modification relationships and polarity shifters were used. The BoosTexter AdaBoost.HM (Schapire and Singer, 2000) achieved 75.9% accuracy on the former tasks and 65.7% on the latter.

When the units of classification, whether words, phrases or sentences, are present in a sequence, identification of sentiment can also be viewed as a tagging task. Breck et al. (2007) used a simple tagging scheme which tagged each term in a sentence as either being ‘in’ a polar expression (/I), or ‘out’ of it (/O). They used various lexical, syntactic and dictionary-based features to identify both /I and /O types of subjective expressions. The MPQA data corpus of 535 newswire articles was used, 135 of which were put aside for parameter tuning and the rest were kept for evaluation. Using 10-fold cross validation and CRFs, the system achieved an F-measure score 70.6%.

Supervised learning with dependency trees was also used by Joshi and Penstein-Rose (2009), who worked on solving the problem of identifying opinions from product reviews. Their method was to transform syntactic dependency relation triplets into features for classification. The motivation was to capture the general relationship between opinionated phrases by 'backing off' to the head word in the triplet. For instance, consider the phrases a great camera and a great mp3 player with the relations {amod, camera, great} and {amod,player,great}. Here, backing off the head words (camera and player) to their POS tags results in a more generalised form {amod,NN,great}, which makes a better indicator for opinion extraction. A collection of 2,200 reviews from the extended version of the Amazon.com/CNet.com product review corpus3 was used, 1,053 of which were subjective. With 11-fold cross-validation in an SVM learner, their method of backing off the head word to the POS achieved approximately 68% accuracy. For obtaining a balanced corpus, 700 documents of each label were selected. N-grams, part-of-speech (POS) tags and their combinations were used as features, and three-fold cross validation was used. Their best system achieved an accuracy of 82.9% when using the unigram presence feature set with SVMs. It should be noted that the corpus used in this work was balanced artificially, thus avoiding the problem of data sparsity for under-represented classes.

Supervised learning can also be performed using multiple classifiers, particularly if the labelling scheme allows for hierarchical relations. As described earlier, one example of this is the work by Pang and Lee (2004). They represented sentences in the given document as graph nodes and calculated the minimal cut on that graph to identify the subjective sentences. Afterwards, standard machine learning classification algorithms (NB and SVMs) were applied only on the extracted subjective sentences to predict their polarity. On a balanced polarity corpus of 2,000 reviews, the minimum-cut framework resulted in an accuracy of 86.4%, which represents a statistically significant improvement in polarity classification accuracy over previous attempts.

In this project, we will also use supervised learning algorithms to classify the citations.

1.3 Unsupervised learning methods

Unsupervised learning is the training of machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. Here the task of machine is to group unsorted information according to similarities, patterns and differences without any prior training of data.

Unlike supervised learning, no teacher is provided that means no training will be given to the machine. Therefore, machine is restricted to find the hidden structure in unlabeled data by our-self.

subjective content was concentrated. Using the AltaVista search engine, this method achieved an overall accuracy of 65%.

Another way to increase the number of items in the lexicon is to use bootstrapping. The general approach is to start with a limited number of polar phrases in the lexicon and to extract similar phrases from unlabelled data. These extracted phrases are then added to the polar lexicon and the process is repeated until a stopping criterion is reached. Riloff and Wiebe's (2003) bootstrapping algorithm learns linguistically rich patterns for subjective expressions. For example, the pattern <subj> was satisfied will match all sentences with the passive form of the verb satisfied. High precision classifiers trained on known subjective vocabulary were used to automatically identify objective and subjective sentences in unannotated text. The labelled sentences from these classifiers were fed to an extraction pattern learner. The remaining unlabelled sentences were filtered through a separate pattern-based subjective sentence classifier that used the extraction pattern previously learned. To close the bootstrap loop, the output of the pattern-based classifier was returned back to the extraction pattern learner and the extracted patterns were used again in the initial high-precision classifiers. This system was used on a balanced corpus of roughly 34,000 sentences and achieved a precision of 0.902 and a recall of 0.401.

Exploring more lexical features in a later work, Wiebe and Riloff (2005) developed a Naive Bayes (NB) classifier using data extracted by a pattern learner. This pattern learner was seeded with known subjective data. Additional features for this NB classifier included strong and weak subjective clues from a pre-existing rule-based system, POS tags, pronouns, modals, adjectives, cardinal number data and adjectives. The classifier was used to classify the unlabelled text corpus, and the most confidently classified sentences were added to the training set for another cycle. They trained the system for two cycles. Using test corpus of 9,289 sentences, 5,104 of which were subjective, they reported up to 0.863 subjective recall with a subjective precision of 0.713 (corresponding to 73.4% accuracy).

While the bootstrapping methods described above determine the similarity between phrases statistically, manually compiled dictionaries and thesauruses can also act as a source of additional lexical items. One such resource is WordNet (Miller, 1995), a large lexical dictionary which provides sets of nouns, verbs, adjectives and adverbs. Each set in WordNet corresponds to a word sense, not a word string. Each set represents a unique concept and members of each set, each of which is a sense, are synonyms of each other.