A Hybrid Approach for Automated Document-level Sentiment Classification (Proposal)

Presented by:
Sara A. Morsy

Supervisor:
Dr. Ahmed Rafea
Overview

• Introduction & Background
• Approaches
• Literature Review
• Problem Statement & Motivation
• Proposed Approach
• Experimentation & Evaluation
Introduction & Background
What is Sentiment Classification?

• Aka Opinion Mining, Sentiment Extraction/Analysis, or Review Mining.
• It is the area of research that attempts to identify the opinion/sentiment that a person may hold towards an object.
• It is a broad area of computational linguistics, natural language processing, text mining and machine learning.
Sentiment Classification Tasks

- At the document-level:
  - Classify a whole document as positive, negative or neutral.

- At the sentence-level:
  - Classify a sentence as subjective or objective and then identify its sentiment as positive, negative or neutral.

- At the feature-level:
  - Classify the opinions on specific features in a single review as positive, negative or neutral.
Applications

• “What other people think” is an important piece of information for most of us:
  ▫ **Individual Interest:**
    • People are interested in others’ opinions when purchasing a product/service or finding opinions on political topics.
  ▫ **Market Intelligence:**
    • Companies are interested in categorizing positive and negative reviews about their products.
  ▫ **Political Interest:**
    • Government intelligence systems seek political opinions expressed online.

• Opinion Search:
  ▫ Not supported by current search engines
• Recommendation & Summarization Systems
Resources and Tools for Sentiment Analysis

Lexicons

Annotated Corpora

Tools
What is a Sentiment Lexicon?

• **Find** relevant words, phrases, and patterns that can be used to express subjectivity
  ▫ **Words**: adjectives, verbs, adverbs and nouns
  ▫ **Phrases containing adjectives and adverbs**
  ▫ **Lexico-syntactic patterns**
    - e.g. expense of <np>: at the expense of the world’s security and stability

• **Determine** the polarity of subjective expressions
Annotated Corpora

- An annotated corpus is needed to:
  - Understand the problem.
  - Create training data and gold standards.
- Annotation is done manually by analyzing the corpus documents and individual sentences and labeling them to their corresponding sentiment (positive, negative or neutral).
Tools

• An online dictionary to search for synonyms and antonyms, e.g. WordNet
• Machine learning classifiers that use text-classification algorithms:
  ▫ Support Vector Machines (SVM)
  ▫ Naïve Bayes (NB)
• Part-Of-Speech (POS) tagger
• Stemmer/Lemmatizer
Challenges

• Subjectivity detection & polarity classification:
  ▫ It can’t be done with just a set of subjective keywords!
  ▫ Context-sensitive
    • This camera is great. (+ve)
    • A great amount of money was spent for promoting this camera. (neutral)
    • If you think this is a great camera, well think again, because ... (-ve)
    • This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it can’t hold up. (-ve)
Challenges (cont’d)

- **Domain-dependent**
  - “Go read the book” can indicate +ve sentiment for book reviews but –ve sentiment for movie reviews.
  - “Unpredictable”: +ve for movie reviews, -ve for car’s steering.

- **Feature-dependent**
  - “long”: - This camera has a long battery life (+ve), vs - The lens of this camera takes a long time to focus (-ve).

- **Topic-sentiment interaction:**
  - “Walmart reports that the profits rose”
    - would be a +ve sentiment if the document is talking about Walmart.
    - would be a -ve sentiment if the document is talking about Target.
Approaches
Document-level Sentiment Classification

- Machine Learning
- Semantic Orientation
- Hybrid Approaches
  - Manually
  - Corpus-based
  - Dictionary-based
1- Machine Learning (ML) Approach

- A classifier is trained using annotated corpora.
- Features used:
  - **Syntactic Features**: e.g. POS tags, n-grams, punctuation
  - **Stylistic Features**:
    - **Lexical Features**: e.g. character- or word-based statistical measures of word variation
    - **Structural Features**: e.g. number of paragraphs
- **It uses text-classification algorithms, such as**:
  - **Support Vector Machines** (best performance)
  - **Naïve Bayes**
  - **AdaBoost**
2- Semantic Orientation (SO) Approach

• A sentiment lexicon is built assigning polarity to sentiment-bearing words/phrases.
• Manually:
  ▫ Labor-intensive task.
  ▫ Done with the other techniques.
• Corpus-based:
  ▫ It finds co-occurrence patterns of words to determine their sentiment polarity.
  ▫ It requires a large corpus.
  ▫ It can find words with domain-specific orientation.
• Dictionary-based:
  ▫ It uses a small seed list in a bootstrapping process to search for synonyms and antonyms in a dictionary to determine their sentiment polarity.
  ▫ It can find a lot of words.
## Pros and Cons of Approaches

<table>
<thead>
<tr>
<th></th>
<th>ML Approach</th>
<th>SO Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Advantages</strong></td>
<td>Performs better on a single domain</td>
<td>Performs better across different domains</td>
</tr>
<tr>
<td></td>
<td>Contextual and domain-specific polarity</td>
<td>Does not require labeled data</td>
</tr>
<tr>
<td><strong>Disadvantages</strong></td>
<td>Requires a large annotated corpus</td>
<td>Prior polarity (no domain-specific polarity)</td>
</tr>
<tr>
<td></td>
<td>Does not take linguistic context into account, e.g. negation &amp; intensification</td>
<td>Needs a manual check for the words and their corresponding polarity and inter-annotator agreement</td>
</tr>
<tr>
<td></td>
<td>Negative classification bias</td>
<td>Positive classification bias</td>
</tr>
</tbody>
</table>
Classification Bias of ML and SO Approaches

• Negative feelings are usually expressed using both positive and negative words, e.g. This car is not good.
• Positive feelings are usually expressed using positive words only.
• The SO approach has a positive classification bias:
  ▫ Since polarities of words are known in advance and positive words sometimes predominate even in negative documents.
• The ML approach has a negative classification bias:
  ▫ Since polarities of words are learnt automatically so it is easier for the classifier to learn negative expressions.
3) Hybrid Approaches

- Many researchers tried combining both ML and SO approaches to make use of their benefits.
- Examples include:
  - Adding sentiment-bearing words with their SO as features for the classifier.
  - Classifying documents with a sentiment lexicon and applying the high-confidence classified set to a classifier as training data.
Literature Review
1) Sentiment Analysis in Multiple Languages

- The authors used both syntactic and stylistic features on an SVM classifier on movie reviews as well as English and Arabic web forums of extremist/hate groups.
- They developed a feature selection algorithm, called Entropy Weighted Genetic Algorithm.
- Using 10-fold cross-validation, they achieved the highest accuracy so far:
  - 91.7% for movie reviews, and
  - 92.8% and 93.6% for English and Arabic web forums, respectively.
2) Lexicon-based Methods for Sentiment Analysis

- The authors developed a sentiment lexicon manually, having separate dictionaries for:
  - Adjectives – nouns – verbs – adverbs – intensifiers

- They also developed a list of:
  - Negators – irrealsis markers (modals, conditional markers (e.g. if), negative polarity items (e.g. any and anything), questions, and words enclosed in quotes)

- Besides, they implemented:
  - Text-level features (e.g. frequency of unigrams) – weighting techniques – multiple cut-offs
2) Lexicon-based Methods for Sentiment Analysis, cont’d

- They achieved high accuracy among different review domains, with overall 78.74% accuracy.
- They achieved accuracy between 62.17-88.98% across different domains, e.g. MPQA, news, blogs and headlines.
- They also outperformed other sentiment lexicons, such as:
3) A Lexicon-Enhanced Method for Sentiment Classification

- The authors used a hybrid approach by adding sentiment-bearing words from SentiWordNet 3.0 as features to an SVM classifier, beside syntactic and stylistic features.
- They applied Information Gain heuristic as a feature selection method.
- Using 10-fold cross-validation, they achieved between 78.85-84.15% accuracy among different review domains.
4) SELC: A Self-Supervised Model for Sentiment Classification

• The authors developed a 2-phase model:
  ▫ 1st phase is lexicon-based:
    • They used a sentiment lexicon and negation word list to classify a set of documents.
  ▫ 2nd phase is corpus-based:
    • They applied the high-confidence classified set as training data for an SVM classifier and classified the uncertain set.
  ▫ Then, results from both phases were integrated to remove the classification bias of both approaches.

• They achieved an overall F1-score of 89.35% across different Chinese review domains.
Flow Chart of the SELC Model

Phase 1

- Input Reviews
  - Initiation Step: Iteration Process: Step 1-4
    - Classified Set
      - A
    - Uncertain Set
      - B
      - Uncertain Set Processing
        - Result of Uncertain Set

For Training
- SVM-HowNet Classifier
  - Result of SVM-HowNet on Uncertain Set
  - Integration Process
    - Result of Integration Process on Uncertain Set
Problem Statement & Motivation

• There is currently no automated domain-independent sentiment classification tool with high accuracy that does not need a manually-annotated corpus.

• Such a tool is needed for opinion search, recommendation, summarization and mining of the increasingly web opinionated content.
Proposed Approach

- Use an efficient sentiment lexicon (verbs, adverbs, adjectives, nouns, and intensifiers) and a negation word list and irrealis markers to classify the documents and apply the high-confidence ones as training data for an SVM classifier.
- Build an SVM classifier with selected syntactic, stylistic and sentiment features.
- Integrate the results of both the sentiment lexicon and the classifier.
- Compare these results with the base-line results.
Experimentation & Evaluation

- Datasets that will be used:
  - The Polarity Dataset (2,000 movie review texts, 1,000 positive and 1,000 negative) provided by Pang and Lee 2004.
  - 400 review texts on: books, cars, computers, cookware, hotels, movies, music, and phones (25 positive and 25 negative reviews in each domain), obtained from Epinions.com (Taboada et al. 2011).
  - A dataset of blogs collected about Jan 25 Revolution.
- Different sentiment, syntactic and stylistic features with different feature selection algorithms will be compared to apply the most efficient ones to the classifier.
Experimentation & Evaluation (cont’d)

• Tools used:
  ▫ Weka Software (SVM classifier)
  ▫ Stanford POS tagger
  ▫ MorphAdorner English Stemmer/Lemmatizer
  ▫ A sentiment lexicon

• Evaluation metrics that will be used:
  ▫ Precision
  ▫ Recall
  ▫ Accuracy
  ▫ $F_1$-score
References

References (cont.)

Thanks!