# Identifying and Analyzing Judgment Opinions

**Soo-Min Kim and Eduard Hovy**
USC Information Sciences Institute
4676 Admiralty Way, Marina del Rey, CA 90292
{skim, hovy}@ISI.EDU

## Abstract

In this paper, we introduce a methodology for analyzing judgment opinions. We define a judgment opinion as consisting of a valence, a holder, and a topic. We decompose the task of opinion analysis into four parts: 1) recognizing the opinion; 2) identifying the valence; 3) identifying the holder; and 4) identifying the topic. In this paper, we address the first three parts and evaluate our methodology using both intrinsic and extrinsic measures.

## 1 Introduction

Recently, many researchers and companies have explored the area of opinion detection and analysis. With the increased immersion of Internet users has come a proliferation of opinions available on the web. Not only do we read more opinions from the web, such as in daily news editorials, but also we post more opinions through mechanisms such as governmental web sites, product review sites, news group message boards and personal blogs. This phenomenon has opened the door for massive opinion collection, which has potential impact on various applications such as public opinion monitoring and product review summary systems.

Although in its infancy, many researchers have worked in various facets of opinion analysis. Pang et al. (2002) and Turney (2002) classified sentiment polarity of reviews at the document level. Wiebe et al. (1999) classified sentence level subjectivity using syntactic classes such as adjectives, pronouns and modal verbs as features. Riloff and Wiebe (2003) extracted subjective expressions from sentences using a bootstrapping pattern learning process. Yu and Hatzivassiloglou (2003) identified the polarity of opinion sentences using semantically oriented words. These techniques were applied and examined in different domains, such as customer reviews (Hu and Liu 2004) and news articles[1]. These researchers use lists of opinion-bearing clue words and phrases, and then apply various additional techniques and refinements.

Along with many opinion researchers, we participated in a large pilot study, sponsored by NIST, which concluded that it is very difficult to define what an opinion is in general. Moreover, an expression that is considered as an opinion in one domain might not be an opinion in another. For example, the statement "The screen is very big" might be a positive review for a wide screen desktop review, but it could be a mere fact in general newspaper text. This implies that it is hard to apply opinion bearing words collected from one domain to an application for another domain. One might therefore need to collect opinion clues within individual domains. In case we cannot simply find training data from existing sources, such as news article analysis, we need to manually annotate data first.

Most opinions are of two kinds: 1) beliefs about the world, with values such as true, false, possible, unlikely, etc.; and 2) judgments about the world, with values such as good, bad, neutral, wise, foolish, virtuous, etc. Statements like "I believe that he is smart" and "Stock prices will rise soon" are examples of beliefs whereas "I like the new policy on social security" and "Unfortunately this really was his year: despite a stagnant economy, he still won his re-election" are examples of judgment opinions. However, judgment opinions and beliefs are not necessarily mutually exclusive. For example, "I think it is an outrage" or "I believe that he is smart" carry both a belief and a judgment.

In the NIST pilot study, it was apparent that human annotators often disagreed on whether a belief statement was or was not an opinion. However, high annotator agreement was seen on judg-

[1] TREC novelty track 2003 and 2004

ment opinions. In this paper, we therefore focus our analysis on judgment opinions only. We hope that future work yields a more precise definition of belief opinions on which human annotators can agree.

We define a judgment opinion as consisting of three elements: a valence, a holder, and a topic. The *valence*, which applies specifically to judgment opinions and not beliefs, is the value of the judgment. In our framework, we consider the following valences: positive, negative, and neutral. The *holder* of an opinion is the person, organization or group whose opinion is expressed. Finally, the *topic* is the event or entity about which the opinion is held.

In previous work, Choi et al. (2005) identify opinion holders (sources) using Conditional Random Fields (CRF) and extraction patterns. They define the opinion holder identification problem as a sequence tagging task: given a sequence of words ($x_1 x_2 \cdots x_n$) in a sentence, they generate a sequence of labels ($y_1 y_2 \cdots y_n$) indicating whether the word is a holder or not. However, there are many cases where multiple opinions are expressed in a sentence each with its own holder. In those cases, finding opinion holders for each individual expression is necessary. In the corpus they used, 48.5% of the sentences which contain an opinion have more than one opinion expression with multiple opinion holders. This implies that multiple opinion expressions in a sentence occur significantly often. A major challenge of our work is therefore not only to focus on sentence with only one opinion, but also to identify opinion holders when there is more than one opinion expressed in a sentence. For example, consider the sentence "*In relation to Bush's axis of evil remarks, the German Foreign Minister also said, Allies are not satellites, and the French Foreign Minister caustically criticized that the United States' unilateral, simplistic worldview poses a new threat to the world*". Here, "*the German Foreign Minister*" should be the holder for the opinion "*Allies are not satellites*" and "*the French Foreign Minister*" should be the holder for "*caustically criticized*".

In this paper, we introduce a methodology for analyzing judgment opinions. We decompose the task into four parts: 1) recognizing the opinion; 2) identifying the valence; 3) identifying the holder; and 4) identifying the topic. For the purposes of this paper, we address the first three parts and leave the last for future work. Opinions can be extracted from various granularities such as a word, a sentence, a text, or even multiple texts. Each is important, but we focus our attention on word-level opinion detection (Section 2.1) and the detection of opinions in short emails (Section 3). We evaluate our methodology using intrinsic and extrinsic measures.

The remainder of the paper is organized as follows. In the next section, we describe our methodology addressing the three steps described above, and in Section 4 we present our experimental results. We conclude with a discussion of future work.

## 2 Analysis of Judgment Opinions

In this section, we first describe our methodology for detecting opinion bearing words and for identifying their valence, which is described in Section 2.1. Then, in Section 2.2, we describe our algorithm for identifying opinion holders. In Section 3, we show how to use our methodology for detecting opinions in short emails.

### 2.1 Detecting Opinion-Bearing Words and Identifying Valence

We introduce an algorithm to classify a word as being positive, negative, or neutral classes. This classifier can be used for any set of words of interest and the resulting words with their valence tags can help in developing new applications such as a public opinion monitoring system. We define an *opinion-bearing word* as a word that carries a positive or negative sentiment directly such as "good", "bad", "foolish", "virtuous", etc. In other words, this is the smallest unit of opinion that can thereafter be used as a clue for sentence-level or text-level opinion detection.

We treat word sentiment classification into Positive, Negative, and Neutral as a three-way classification problem instead of a two-way classification problem of Positive and Negative. By adding the third class, Neutral, we can prevent the classifier from assigning either positive or negative sentiment to weak opinion-bearing words. For example, the word "central" that Hatzivassiloglou and McKeown (1997) included as a positive adjective is not classified as positive in our system. Instead

we mark it as "neutral" since it is a weak clue for an opinion. If an unknown word has a strong relationship with the neutral class, we can therefore classify it as neutral even if it has some small connotation of Positive or Negative as well.

**Approach**: We built a word sentiment classifier using WordNet and three sets of positive, negative, and neutral words tagged by hand. Our insight is that synonyms of positive words tend to have positive sentiment. We expanded those manually selected seed words of each sentiment class by collecting synonyms from WordNet. However, we cannot simply assume that all the synonyms of positive words are positive since most words could have synonym relationships with all three sentiment classes. This requires us to calculate the closeness of a given word to each category and determine the most probable class. The following formula describes our model for determining the category of a word:

$$\arg\max_c P(c \mid w) \cong \arg\max_c P(c \mid syn_1, syn_2 \ldots . syn_n) \qquad (1)$$

where $c$ is a category (Positive, Negative, or Neutral) and $w$ is a given word; $syn_n$ is a WordNet synonym of the word $w$. We calculate this closeness as follows;

$$\begin{aligned} &\arg\max_c P(c \mid w) = \arg\max_c P(c) P(w \mid c) \\ &= \arg\max_c P(c) P(syn_1 syn_2 syn_3 \ldots syn_n \mid c) \\ &= \arg\max_c P(c) \prod_{k=1}^{m} P(f_k \mid c)^{count(f_k, synset(w))} \qquad (2) \end{aligned}$$

where $f_k$ is the $k^{th}$ feature of class $c$ which is also a member of the synonym set of the given word $w$. $count(f_k, synset(w))$ is the total number of occurrences of the word feature $f_k$ in the synonym set of word $w$. In section 4.1, we describe our manually annotated dataset which we used for seed words and for our evaluation.

## 2.2 Identifying Opinion Holders

Despite successes in identifying opinion expressions and subjective words/phrases (See Section 1), there has been less achievement on the factors closely related to subjectivity and polarity, such as identifying the opinion holder. However, our research indicates that *without* this information, it is difficult, if not impossible, to define 'opinion' accurately enough to obtain reasonable inter-annotator agreement. Since these factors co-occur and mutually reinforce each other, the question "Who is the holder of this opinion?" is as important as "Is this an opinion?" or "What kind of opinion is expressed here?".

| | |
|---|---|
| Sentence | Iraqi Vice President Taha Yassin Ramadan, responding to Bush's 'axis of evil' remark, said the U.S. government 'is the source of evil' in the world. |
| Expressive subjectivity | the U.S. government 'is the source of evil' in the world |
| Strength | Extreme |
| Source | Iraqi Vice President Taha Yassin Ramadan |

Table 1: Annotation example

In this section, we describe the automated identification for opinion holders. We define an opinion holder as an entity (person, organization, country, or special group of people) who expresses explicitly or implicitly the opinion contained in the sentence.

Previous work that is related to opinion holder identification is (Bethard et al. 2004) who identify opinion propositions and holders. However, their opinion is restricted to propositional opinion and mostly to verbs. Another related work is (Choi et al. 2005) who use the MPQA corpus[2] to learn patterns of opinion sources using a graphical model and extraction pattern learning. However, they have a different task definition from ours. They define the task as identifying opinion sources (holders) given a sentence, whereas we define it as identifying opinion sources given an opinion expression in a sentence. We discussed their work in Section 1.

**Data**: As training data, we used the MPQA corpus (Wilson and Wiebe, 2003), which contains news articles manually annotated by 5 trained annotators. They annotated 10657 sentences from 535 documents, in four different aspects: *agent, expressive-subjectivity, on,* and *inside*. *Expressive-subjectivity* marks words and phrases that indirectly express a *private state* that is defined as a term for opinions, evaluations, emotions, and speculations. The *on* annotation is used to mark speech events and direct expressions of private states. As for the holder, we use the agent of the selected private states or speech events. While there are many possible ways to define what opinion means, intuitively, given an opinion, it is clear what the opinion holder means. Table 1 shows an example of the annotation. In this example, we consider the expression "the U.S. government 'is the source of evil' in the world" with an expres-

[2] http://www.cs.pitt.edu/~wiebe/pubs/ardasummer02/

sive-subjectivity tag as an opinion of the holder "Iraqi Vice President Taha Yassin Ramadan".

**Approach:** Since more than one opinion may be expressed in a sentence, we have to find an opinion holder for each opinion expression. For example, in a sentence "A thinks B's criticism of T is wrong", B is the holder of "the criticism of T", whereas A is the person who has an opinion that B's criticism is wrong. Therefore, we define our task as finding an opinion holder, given an opinion expression. Our earlier work focused on identifying opinion expressions within text. We employ that system in tandem with the one described here.

To learn opinion holders automatically, we use a Maximum Entropy model. Maximum Entropy models implement the intuition that the best model is the one that is consistent with the set of constraints imposed by the evidence but otherwise is as uniform as possible (Berger et al. 1996). There are two ways to model the problem with ME: classification and ranking. Classification allocates each holder candidate to one of a set of predefined classes while ranking selects a single candidate as answer. This means that classification modeling[3] can select many candidates as answers as long as they are marked as true, and does not select any candidate if every one is marked as false. In contrast, ranking always selects the most probable candidate as an answer, which suits our task better. Our earlier experiments showed poor performance with classification modeling, an experience also reported for Question Answering (Ravichandran et al. 2003).

We modeled the problem to choose the most probable candidate that maximizes a given conditional probability distribution, given a set of holder candidates $\{h_1 h_2 \ldots h_N\}$ and opinion expression $e$. The conditional probability $P(h|\{h_1 h_2 \ldots h_N\}, e)$ can be calculated based on $K$ feature functions $f_k(h, \{h_1 h_2 \ldots h_N\}, e)$. We write a decision rule for the ranking as follows:

$$h = \arg\max_h [P(h | \{h_1 h_2 \ldots h_N\}, e)]$$

$$= \arg\max_h [\sum_{k=1}^{K} \lambda_k f_k(h, \{h_1 h_2 \ldots h_N\}, e)]$$

Each $\lambda_k$ is a model parameter indicating the weight of its feature function.

[3] In our task, there are two classes: holder and non-holder.

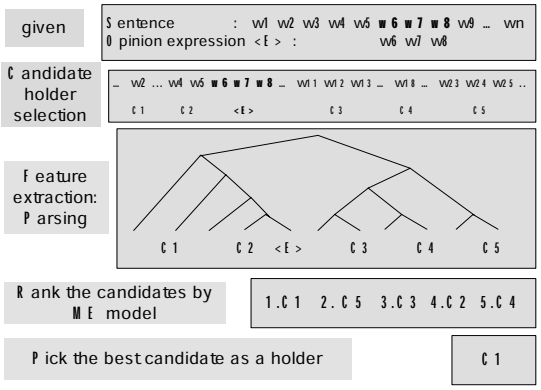

Figure 1: Overall system architecture

Figure 1 illustrates our holder identification system. First, the system generates all possible holder candidates, given a sentence and an opinion expression <E>. After parsing the sentence, it extracts features such as the syntactic path information between each candidate <H> and the expression <E> and a distance between <H> and <E>. Then it ranks holder candidates according to the score obtained by the ME ranking model. Finally the system picks the candidate with the highest score. Below, we describe in turn how to select holder candidates and how to select features for the training model.

**Holder Candidate Selection:** Intuitively, one would expect most opinion holders to be named entities (PERSON or ORGANIZATION)[4]. However, other common noun phrases can often be opinion holders, such as "the leader", "three nations", and "the Arab and Islamic world". Sometimes, pronouns like *he, she,* and *they* that refer to a PERSON, or *it* that refers to an ORGANIZATION or country, can be an opinion holder. In our study, we consider all noun phrases, including common noun phrases, named entities, and pronouns, as holder candidates.

**Feature Selection:** Our hypothesis is that there exists a structural relation between a holder <H> and an expression <E> that can help to identify opinion holders. This relation may be represented by lexical-level patterns between <H> and <E>, but anchoring on surface words might run into the data sparseness problem. For example, if we see the lexical pattern "<H> recently criticized <E>" in the training data, it is impossible to match the expression "<H> yesterday condemned <E>". These, however, have the same syntactic features in our

[4] We use BBN's named entity tagger IdentiFinder to collect named entities.

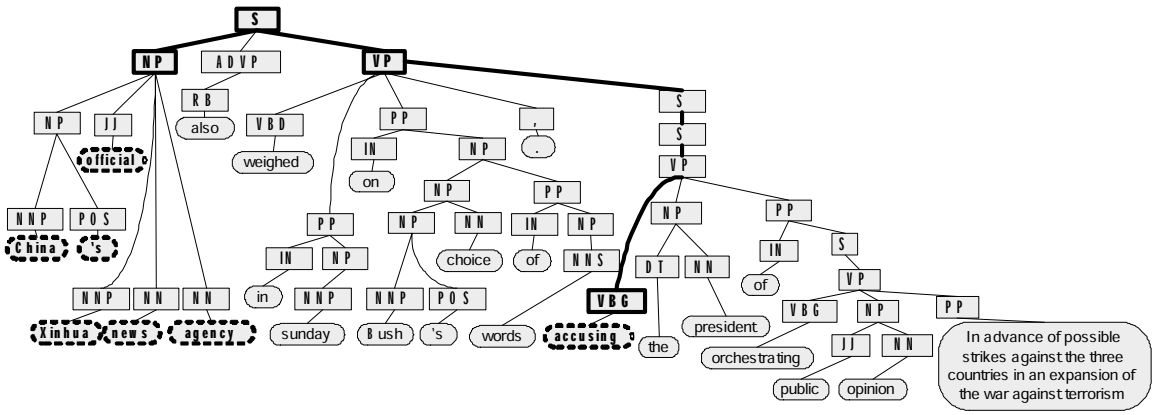


Figure 2: A parsing example

model. We therefore selected structural features from a deep parse, using the Charniak parser.

After parsing the sentence, we search for the *lowest common parent node* of the words in <H> and <E> respectively (<H> and <E> are mostly expressed with multiple words). A lowest common parent node is a non-terminal node in a parse tree that covers all the words in <H> and <E>. Figure 2 shows a parsed example of a sentence with the holder "China's official Xinhua news agency" and the opinion expression "accusing". In this example, the lowest common parent of words in <H> is the bold NP and the lowest common parent of <E> is the bold VBG. We name these nodes *Hhead* and *Ehead* respectively. After finding these nodes, we label them by subscript (e.g., $NP_H$ and $VBG_E$) to indicate they cover <H> and <E>. In order to see how Hhead and Ehead are related to each other in the parse tree, we define another node, *HEhead,* which covers both Hhead and Ehead. In the example, HEhead is S at the top of the parse tree since it covers both $NP_H$ and $VBG_E$. We also label S by subscript as $S_{HE}$.

To express tree structure for ME training, we extract path information between <H> and <E>. In the example, the complete path from Hhead to Ehead is "<H> NP S VP S S VP VBG <E>". However, representing each complete path as a single feature produces so many different paths with low frequencies that the ME system would learn poorly. Therefore, we split the path into three parts: *HEpath, Hpath an Epath*. HEpath is defined as a path from HEhead to its left and right child nodes that are also parents of Hhead and Ehead. Hpath is a path from Hhead and one of its ancestor nodes that is a child of HEhead. Similarly, Epath is defined as a path from Ehead to one of its ancestors that is also a child of HEhead. With this splitting, the system can work when any of HEpath, Hpath or Epath appeared in the training data, even if the entire path from <H> to <E> is unseen. Table 2 summarizes these concepts with two holder candidate examples in the parse tree of Figure 2.

We also include two non-structural features. The first is the type of the candidate, with values NP, PERSON, ORGANIZATION, and LOCATION. The second feature is the distance between <H> and <E>, counted in parse tree words. This is motivated by the intuition that holder candidates tend to lie closer to their opinion expression. All features are listed in Table 3. We describe the performance of the system in Section 4.

| | Candidate 1 | Candidate 2 |
|---|---|---|
| | China's official Xinuhua news agency | Bush |
| Hhead | $NP_H$ | $NNP_H$ |
| Ehead | $VBG_E$ | $VBG_E$ |
| HEhead | $S_{HE}$ | $VP_{HE}$ |
| Hpath | $NP_H$ | $NNP_H$ $NP_H$ $NP_H$ $NP_H$ $PP_H$ |
| Epath | $VBG_E$ $VP_E$ $S_E$ $S_E$ $VP_E$ | $VBG_E$ $VP_E$ $S_E$ $S_E$ |
| HEpath | $S_{HE}$ $NP_H$ $VP_E$ | $VP_{HE}$ $PP_H$ $S_E$ |

Table 2: Heads and paths for the Figure 2 example

| Features | Description |
|---|---|
| F1 | Type of <H> |
| F2 | HEpath |
| F3 | Hpath |
| F4 | Epath |
| F5 | Distance between <H> and <E> |

Table 3: Features for ME training

| Model 1 |
| --- |
| · Translate a German email to English<br>· Apply English opinion-bearing words |
| **Model 2** |
| · Translate English opinion-bearing words to German<br>· Analyze a German email using the German opinion-bearing words. |

Table 4: Two models of German Email opinion analysis system

## 3 Applying our Methodology to German Emails

In this section, we describe a German email analysis system into which we included the opinion-bearing words from Section 2.1 to detect opinions expressed in emails. This system is part of a collaboration with the EU-funded project QUALEG (Quality of Service and Legitimacy in eGovernment) which aims at enabling local governments to manage their policies in a transparent and trustable way[5]. For this purpose, local governments should be able to measure the performance of the services they offer, by assessing the satisfaction of its citizens. This need makes a system that can monitor and analyze citizens' emails essential. The goal of our system is to classify emails as neutral or as bearing a positive or negative opinion.

To generate opinion bearing words, we ran the word sentiment classifier from Section 2.1 on 8011 verbs to classify them into 807 positive, 785 negative, and 6149 neutral. For 19748 adjectives, the system classified them into 3254 positive, 303 negative, and 16191 neutral. Since our opinion-bearing words are in English and our target system is in German, we also applied a statistical word alignment technique, GIZA++ [6] (Och and Ney 2000). Running it on version two of the European Parliament corpus, we obtained statistics for 678,340 German-English word pairs and 577,362 English-German word pairs. Obtaining these two lists of translation pairs allows us to convert English words to German, and German to English, without a full document translation system. To utilize our English opinion-bearing words in a German opinion analysis system, we developed two models, outlined in Table 4, each of which is triggered at different points in the system.

In both models, however, we still need to decide how to apply opinion-bearing words as clues to determine the sentiment of a whole email. Our previous work on sentence level sentiment classification (Kim and Hovy, 2004) shows that the presence of any negative words is a reasonable indication of a negative sentence. Since our emails are mostly short (the average number of words in each email is 19.2) and we avoided collecting weak negative opinion clue words, we hypothesize that our previous sentence sentiment classification study works on the email sentiment analysis. This implies that an email is negative if it contains more than certain number of strong negative words. We tune this parameter using our training data. Conversely, if an email contains mostly positive opinion-bearing words, we classify it as a positive email. We assign neutral if an email does not contain any strong opinion-bearing words.

Manually annotated email data was provided by our joint research site. This data contains 71 emails from citizens regarding a German festival. 26 of them contained negative complaints, for example, the lack of parking space, and 24 of them were positive with complimentary comments to the organization. The rest of them were marked as "questions" such as how to buy festival tickets, "only text" of simple comments, "fuzzy", and "difficult". So, we carried system experiments on positive and negative emails with precision and recall. We report system results in Section 4.

## 4 Experiment Results

In this section, we evaluate the three systems described in Sections 2 and 3: detecting opinion-bearing words and identifying valence, identifying opinion holders, and the German email opinion analysis system.

### 4.1 Detecting Opinion-bearing Words

We described a word classification system to detect opinion-bearing words in Section 2.1. To examine its effectiveness, we annotated 2011 verbs and 1860 adjectives, which served as a gold standard[7]. These words were randomly selected from a

[5] http://www.qualeg.eupm.net/my_spip/index.php
[6] http://www.fjoch.com/GIZA++.html

[7] Although nouns and adverbs may also be opinion-bearing, we focus only on verbs and adjectives for this study.

| | Positive | Negative | Neutral | Total |
|---|---|---|---|---|
| **Verb** | 69 | 151 | 1791 | 2011 |
| **Adjective** | 199 | 304 | 1357 | 1860 |

Table 5: Word distribution in our gold standard

| | | Precision | Recall | F-score |
|---|---|---|---|---|
| P | V | 20.5% ± 3.5% | 82.4% ± 7.5% | 32.3% ± 4.6% |
| | A | 32.4% ± 3.8% | 75.5% ± 6.1% | 45.1% ± 4.4% |
| X | V | 97.2% ± 0.6% | 77.6% ± 1.4% | 86.3% ± 0.7% |
| | A | 89.5% ± 1.7% | 67.1% ± 2.7% | 76.6% ± 2.1% |
| N | V | 37.8% ± 4.9% | 76.2% ± 8.0% | 50.1% ± 5.6% |
| | A | 60.0% ± 4.1% | 78.5% ± 4.9% | 67.8% ± 3.8% |

Table 6: Precision, recall, and F-score on word valence categorization for Positive (P), Negative (N) and Neutral (X) verbs (V) and adjectives (A) (with 95% confidence intervals)

collection of 8011 English verbs and 19748 English adjectives. We use training data as seed words for the WordNet expansion part of our algorithm (described in Section 2.1). Table 5 shows the distribution of each semantic class. In both verb and adjective annotation, neutral class has much more words than the positive or negative classes.

We measured the precision, recall, and F-score of our system using 10-fold cross validation. Table 6 shows the results with 95% confidence bounds. Overall (combining positive, neutral and negative), our system achieved 77.7% ± 1.2% accuracy on verbs and 69.1% ± 2.1% accuracy on adjectives. The system has very high precision in the neutral category for both verbs (97.2%) and adjectives (89.5%), which we interpret to mean that our system is really good at filtering non-opinion bearing words. Recall is high in all cases but precision varies; very high for neutral and relatively high for negative but low for positive.

### 4.2 Opinion Holder Identification

We conducted experiments on 2822 <sentence; opinion expression; holder> triples and divided the data set into 10 <training; test> sets for cross validation. For evaluation, we consider to match either fully or partially with the holder marked in the test data. The holder matches *fully* if it is a single entity (e.g., "Bush"). The holder matches *partially* when it is part of the multiple entities that make up the marked holder. For example, given a marked holder "Michel Sidibe, Director of the Country and Regional Support Department of UNAIDS", we consider both "Michel Sidibe" and "Director of the Country and Regional Support Department of UNAIDS" as acceptable answers.

| | **Baseline** | **F5** | **F15** | **F234** | **F12345** |
|---|---|---|---|---|---|
| Top1 | 23.2% | 21.8% | 41.6% | 50.8% | 52.7% |
| Top2 | | 39.7% | 61.9% | 66.3% | 67.9% |
| Top3 | | 52.2% | 72.5% | 77.1% | 77.8% |

Table 7: Opinion holder identification results (excluding pronouns from candidates)

| | **Baseline** | **F5** | **F15** | **F234** | **F12345** |
|---|---|---|---|---|---|
| Top1 | 21.3% | 18.9% | 41.8% | 47.9% | 50.6% |
| Top2 | | 37.9% | 61.6% | 64.8% | 66.7% |
| Top3 | | 51.2% | 72.3% | 75.3% | 76.0% |

Table 8: Opinion holder identification results (All noun phrases as candidates)

Our experiments consist of two parts based on the candidate selection method. Besides the selection method we described in Section 2.2, we also conducted a separate experiment by excluding pronouns from the candidate list. With the second method, the system always produces a non-pronoun holder as an answer. This selection method is useful in some Information Extraction application that only cares non-pronoun holders.

We report accuracy (the percentage of correct answers the system found in the test set) to evaluate our system. We also report how many correct answers were found within the top2 and top3 system answers. Tables 7 and 8 show the system accuracy with and without considering pronouns as alias candidates, respectively. Table 8 mostly shows lower accuracies than Table 7 because test data often has only a non-pronoun entity as a holder and the system picks a pronoun as its answer. Even if the pronoun refers the same entity marked in the test data, the evaluation system counts it as wrong because it does not match the hand annotated holder.

To evaluate the effectiveness of our system, we set the baseline as a system choosing the closest candidate to the expression as a holder without the Maximum Entropy decision. The baseline system had an accuracy of only 21.3% for candidate selection over all noun phrases and 23.2% for candidate selection excluding pronouns.

The results show that detecting opinion holders is a hard problem, but adopting syntactic features (F2, F3, and F4) helps to improve the system. A promising avenue of future work is to investigate the use of semantic features to eliminate noun

phrases such as "cheap energy subsidies" or "possible strikes" from the candidate set before we run our ME model, since they are less likely to be an opinion holder than noun phrases like "three nations" or "Palestine people."

### 4.3 German Emails

For our experiment, we performed 7-fold cross validation on a set of 71 emails. Table 9 shows the average precision, recall, and F-score. Results show that our system identifies negative emails (complaints) better than praise. When we chose a system parameter for the focus, we intended to find negative emails rather than positive emails because officials who receive these emails need to act to solve problems when people complain but they have less need to react to compliments. By highlighting high recall of negative emails, we may misclassify a neutral email as negative but there is also less chance to neglect complaints.

| Category | | Model1 | Model2 |
|---|---|---|---|
| Positive (P) | Precision | 0.72 | 0.55 |
| | Recall | 0.40 | 0.65 |
| | F-score | 0.51 | 0.60 |
| Negative (N) | Precision | 0.55 | 0.61 |
| | Recall | 0.80 | 0.42 |
| | F-score | 0.65 | 0.50 |

Table 9: German email opinion analysis system results

## 5 Conclusion and Future Work

In this paper, we presented a methodology for analyzing judgment opinions, which we define as opinions consisting of a valence, a holder, and a topic. We presented models for recognizing sentences containing judgment opinions, identifying the valence of the opinion, and identifying the holder of the opinion. Remaining is to also finally identify the topic of the opinion. Past tests with human annotators indicate that the accuracy of identifying valence, holder and topic is much increased when all three are being done simultaneously. We plan to investigate a joint model to verify this intuition.

Our past work indicated that, for newspaper texts, it is feasible for annotators to identify judgment opinion sentences and for them to identify their holders and judgment valences. It is encouraging to see that we achieved good results on a new genre – emails sent from citizens to a city council – and in a new language, German.

This paper presents a computational framework for analyzing judgment opinions. Even though these are the most common opinions, it is a pity that the research community remains unable to define *belief* opinions (i.e., those opinions that have values such as true, false, possible, unlikely, etc.) with high enough inter-annotator agreement. Only once we properly define belief opinion will we be capable of building a complete opinion analysis system.

## References


Berger, A, S. Della Pietra, and V. Della Pietra. 1996. A Maximum Entropy Approach to Natural Language. *Computational Linguistics 22(1).*

Bethard, S., H. Yu, A. Thornton, V. Hatzivassiloglou, and D. Jurafsky. 2004. Automatic Extraction of Opinion Propositions and their Holders. *AAAI Spring Symposium on Exploring Attitude and Affect in Text.*

Charniak, E. 2000. A Maximum-Entropy-Inspired Parser. *Proc. of NAACL-2000.*

Choi, Y., C. Cardie, E. Riloff, and S. Patwardhan. 2005. Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns. *Proc. of Human Language Technology Conference/Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP 2005).*

Esuli, A. and F. Sebastiani. 2005. Determining the semantic orientation of terms through gloss classification. *Proc. of CIKM-05, 14th ACM International Conference on Information and Knowledge Management.*

Hatzivassiloglou, V. and McKeown, K. (1997). Predicting the semantic orientation of adjectives. *Proc. 35th Annual Meeting of the Assoc. for Computational Linguistics (ACL-EACL 97.*

Hu, M. and Liu, B. 2004. Mining and summarizing customer reviews. *Proc. of KDD'04. pp.168 - 177*

Och, F.J. 2002. Yet Another MaxEnt Toolkit: YASMET http://wasserstoff.informatik.rwth-aachen.de/Colleag ues/och/

Och, F.J and Ney, H. 2000. Improved statistical alignment models. *Proc. of ACL-2000, pp. 440–447, Hong Kong, China.*

Pang, B, L. Lee, and S. Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proc. of EMNLP 2002.*

Ravichandran, D., E. Hovy, and F.J. Och. 2003. Statistical QA - classifier vs re-ranker: What's the difference? *Proc. of the ACL Workshop on Multilingual Summarization and Question Answering.*

Riloff, E. and J. Wiebe. 2003. Learning Extraction Patterns for Subjective Expressions. *Proc. of EMNLP-03.*

Turney, P. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proc. of the 40th Annual Meeting of the ACL, 417–424.*

Wiebe, J, R. Bruce, and T. O'Hara. 1999. Development and use of a gold standard data set for subjectivity classifications. *Proc. of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99), 246–253.*

Wilson, T. and J. Wiebe. 2003. Annotating Opinions in the World Press. *Proc. of ACL SIGDIAL-03.*

Yu, H. and V. Hatzivassiloglou. 2003. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. *Proc. of EMNLP.*