Automatic Detection of Point of View Differences in Wikipedia

Khalid Al Khatib[1] Hinrich Schütze[1] Cathleen Kantner[2]
(1) Institute for Natural Language Processing
University of Stuttgart
(2) Institute for Social Sciences
University of Stuttgart
khalid@ims.uni-stuttgart.de

ABSTRACT
We investigate differences in point of view (POV) between two objective documents, where one is describing the subject matter in a more positive/negative way than the other, and present an automatic method for detecting such POV differences. We use Amazon Mechanical Turk (AMT) to annotate sentences as positive, negative or neutral based on their POV towards a given target. A statistical classifier is trained to predict the *POV score* of a document, which reflects how positive/negative the document's POV towards its target is. The results of our experiments on a set of articles in the Arabic and English Wikipedias from the *people* category show that our method successfully detects POV differences.

KEYWORDS: sentiment analysis, content analysis, natural language processing.

1 Introduction

In many areas of public discourse, content creators strive for objectivity. Reporters try to report the facts without bias; judges are expected to write opinions that are not influenced by personal views; encyclopedias are committed to what Wikipedia calls a "neutral point of view" (NPOV), defined as "... representing fairly, proportionately, and as far as possible without bias, all significant views that have been published by reliable sources".[1]

Even though objectivity is an important ideal, content creators cannot avoid being influenced by their background and context. There are two main reasons for this (Scheufele, 1999; Habermas, 2006; Littlejohn and Foss, 2010; D'Alessio and Allen, 2000). First, there are always many different non-equivalent ways of conveying a given piece of information. By choosing one vs. the other, the content creator introduces part of his/her point of view (POV) into the discourse. For example, "his wars caused the death of more than a million civilians" (a translation of a sentence in the French Wikipedia) puts Napoleon in a more negative light than "more than a million civilians died in his wars", which in turn is more negative than "more than a million civilians died in the wars fought between him and his enemies". This is so because the chain of causality between Napoleon and people being killed is more explicit in the first sentence than in the third. None of these sentences is a violation of Wikipedia's neutral point of view policy.

The second reason for different objective points of view is that the selection of what information to present is also influenced by background and context. If for space reasons only one of two equally relevant facts about a politician – one positive, one negative – can be added to a newspaper article, then this choice impacts how positive/negative the article is. This is an unavoidable dilemma journalists face on a daily basis. It is usually impossible to include all available information.

We call the difference between two objective documents, where one describes the subject in a more positive/negative way than the other, a *point of view difference* or *POV difference*. This paper develops a method that detects POV differences and quantifies their magnitude.

The automatic identification of POV and POV differences is of high potential for content analysis in the social sciences – which we take to include the humanities in this paper. Early content analysis was motivated by concerns about the declining quality of public debate in modern mass societies and tried to answer empirically questions such as: Do the media live up to their own quality standards of factually accurate and ideologically unbiased reporting? Are no relevant facts omitted? Are all relevant POVs equally represented? (Krippendorff, 2004, 55ff.)

There are also systematic reasons for the widespread empirical investigation of the evaluative positions taken by various speakers in the media. Our social world is permanently produced and reproduced, interpreted and criticized in social interactions of actors with their – often conflicting – intentions, values and reasons for actions (Berger and Luckmann, 1966; Habermas, 1984). This activity leaves traces – interpretable symbols, text and images. Evaluative aspects are almost always of central importance for the social science research question that motivates a content analysis. Typical questions that scholars, readers, the public and practitioners ask are: How favorably are the objects (e.g., social groups, politicians, policies, countries, corporations) perceived? Do different groups of people (e.g., migrants, citizens of different countries) view a certain object differently in a systematic fashion? How can this be explained? Which effect will this have?

[1] http://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view

In most content analysis today, positive and negative presentation of subject are annotated – or coded – *manually*, an expensive and time-consuming method. To maximize intercoder agreement, the coding usually is restricted to explicit evaluative claims although some authors have also looked at stylistic means – such as irony and emotional language – for weakening or strengthening evaluations (Früh, 2011, 241–260). The result of this type of manual analysis is usually an aggregate value of the variable of interest that can then be used to compare groups (e.g., French vs. U.S. newspapers) or to analyze changes over time.

These efforts have not resulted in a widely shared research methodology. This may be due to the problem that there will never be an "objective" standard of what would constitute a pure, neutral way of reporting (Krippendorff, 2004, 55–57). Moreover, the subtle differences in evaluative tone and connotation that are of so much interest to social scientists are difficult to operationalize in a reliable way.

The approach presented in this paper overcomes these two problems of current content analysis. First, we develop a fully automatic method that is suitable for the analysis of large amounts of text and thereby reduces the obstacles that manual analysis and reliance on human coders present.

Second, while we acknowledge that the *absolute* assessment of POV is an important problem, we do not pretend to define an objective neutral standard in this paper. Instead, we cast the problem of assessing POVs as a *relative* problem and thereby avoid the difficulties inherent in attempting to define objective standards.

POV differences are also related to work on sentiment analysis in natural language processing (NLP). In contrast to most prior work in sentiment analysis, we are concerned only with *objective language* in this paper. For example, we do not address the analysis of editorials (which are intentionally opinionated) or of badly written Wikipedia articles (which violate the NPOV principle). The question as to how to automatically assess whether a piece of text written in objective language represents a positive or negative POV of the subject matter and to what extent has not been addressed before.

In addition to defining the problem of POV difference detection and proposing a method for solving it, we also provide an *evaluation gold standard*. It consists of articles and sentences from the Arabic and English Wikipedias that were annotated for POV and for POV differences using a combination of Amazon Mechanical Turk and student annotators. We chose the Arabic and English Wikipedias as the basis for our data set because we found that it contains many POV differences.

This paper is organized as follows. Section 2 introduces and motivates the concept of POV difference. Section 3 presents data acquisition and preparation. We describe our approach to the detection of POV differences in Section 4. Section 5 presents experiments and evaluates results. In Section 6, we discuss our results. Section 7 covers related work. Finally, conclusions and a brief summary of planned future work are given in Section 8.

2 Point of view (POV) differences

As a concrete example for a POV difference consider the French and Spanish Wikipedia articles about Napoleon.[2] Both articles are objective and meet the neutral POV criteria of Wikipedia. However, there is a POV difference between them: the French article is more positive than

[2] Based on the versions available online on 2012-04-01.

the Spanish article. We can find instances of the two types of reasons for POV differences that we discussed above: (i) different ways of describing a certain fact and (ii) different ways of selecting subsets of facts.

An example of different descriptions of the same facts is the phrase “sus agresivas guerras de conquista” ‘his aggressive wars of conquest’ in the Spanish Wikipedia. This amounts to a negative evaluation of Napoleon. Nowhere in the French article are Napoleon’s wars called aggressive. Instead, his readiness to attack and the speed of his campaigns are referred to in more positive words: “offensive immédiate” ‘immediate offensive’, “marche forcée” ‘forced march’, “impressionante de rapidité” ‘impressive for its rapidity’.

An example of a different selection of facts is the number of casualties during the Peninsular war. Only the Spanish article gives an estimate (300,000), which potentially casts a negative light on Napoleon as someone who is responsible for the loss of many lives.

Our goal in this paper is to develop a method that detects and quantifies such POV differences. Our approach is to first train a classifier that detects absolute POV. We then calculate POV difference between two articles as the difference of the absolute POV scores.

3 Data acquisition and preparation

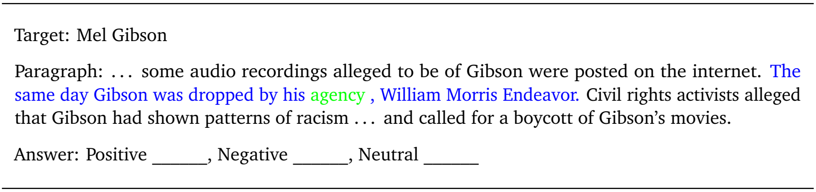
Our approach to estimating POV differences for a pair of documents is to *estimate the absolute POV score* for each of the two articles of the pair and then *calculate the difference*. This approach will be described in detail in Section 4. To build a POV difference detector and evaluate it, we need a gold standard. Our gold standard consists of two parts, one for absolute POV and one for POV differences. While we could limit ourselves to only evaluating our main task, the estimation of POV differences, we decided instead to also evalute the quality of absolute POV scores. In this section, we describe the two gold standards we need for executing this plan: the gold standard for absolute POV scores and the gold standard for POV differences.

Gold standard for absolute POV. We first must decide which unit of text to create the gold standard for. Even though we are interested in the evaluation of entire documents, we do not annotate documents for two reasons. First, most documents will contain a mix of different POVs, so that a single label gives a statistical classifier noisy information. Second, reading and evaluating an entire document takes a long time for an annotator and would make gold standard creation expensive.

On the other hand, our units cannot be too small – e.g., words or phrases – because POV is a complex phenomenon that cannot be judged reliably at such a low level; the sentences about Napoleon in the introduction are examples for this.

Based on this reasoning we choose *the sentence* as our annotation unit. We annotate two sets of 1200 sentences, one for Arabic and one for English. The Arabic (resp. English) set consists of the first 20 sentences of 60 Arabic (resp. English) Wikipedia articles from the category *people*. We selected articles about people that are well known in both Western culture and Arabic culture because they are more likely to have been written by experienced authors and therefore to have a high quality.

The second major decision concerns the classification scheme for POV. We define three *POV classes*: positive, neutral and negative. These classes cover the potential cases of POV. We need a neutral class since many sentences do not contain any information that implies positive or negative POV.



Target: Mel Gibson

Paragraph: . . . some audio recordings alleged to be of Gibson were posted on the internet. The same day Gibson was dropped by his agency , William Morris Endeavor. Civil rights activists alleged that Gibson had shown patterns of racism . . . and called for a boycott of Gibson's movies.

Answer: Positive ______, Negative ______, Neutral ______

Figure 1: Interface of the AMT task.

The final decision concerns the annotators. We use Amazon Mechanical Turk (AMT). AMT has become a standard method for gold standard creation in NLP because annotations are of reasonable quality and comparatively low in cost (cf. (Alonso and Lease, 2011)).

For many objective statements, it is clear which POV class – positive, neutral or negative – applies to them. However, there is a certain subset of statements for which the decision is difficult. The different degrees of explicitness in describing a causal chain from Napoleon's actions to people dying in the introduction are a good example. As the statements become more explicit about the causal relation, at some point the sentence acquires a negative POV, but people differ as to when that point is reached.

Figure 1 shows the interface of the AMT task.[3] We ask non-expert workers to provide annotations for POV, a difficult decision for a subset of sentences. Thus, the design of the HIT (Human Intelligence Task) in AMT is crucial: in order to get acceptable agreement, the AMT task must be simple and easy to understand; definitions must be clear and the annotation interface well-structured.

Definitions of the three POV classes are provided in the instructions: the sentence has a positive (resp. negative, neutral) POV toward the target if it states that the target did something positive (resp. negative, neutral) or is described in a positive (resp. negative, neutral) way. No direct information about the target is also rated as neutral. Four examples from Wikipedia articles are given to help workers understand the task: one for positive, one for negative, and two for the neutral POV class. One neutral example shows a sentence that is directly relevant to the target, but is neither positive nor negative. The other neutral example is a negative sentence that is not relevant to assessing the POV of the article towards the target – e.g., because it talks about historical background that the target is not involved in. Because we found almost no sentences that had a mix of positive/negative elements, we did not explicitly include this case in the instructions.

The instructions are appropriately adapted for Arabic and English. They state that the Arabic (resp. English) task is only for Arabic (resp. English) native speakers. Even though the workers of the Arabic task have to be Arabic native speakers, the language of the instructions is English. All AMT workers know English since the AMT platform has only an English interface; so English as instruction language does not impose any additional restrictions on eligibility.

Each task includes one of the 1200 selected sentences (in blue color), the target that the article the sentence is extracted from is about (top line in Figure 1: "Mel Gibson"), and the surrounding paragraph (in black). To ensure sufficient context, we show to workers the entire paragraph

[3] We have reformatted the output that annotators see for space reasons and better legibility. E.g., we have omitted some text (marked ". . . "). Annotators see the entire paragraph without omissions.

	Arabic	English
agreement – all workers	.30	.48
agreement – majority of workers	.90	.95
Fleiss' κ	.215	.419

	Arabic	English
positive	.435	.47
neutral	.42	.34
negative	.145	.19

Table 1: Absolute POV gold standard: agreement (left); sentence label distribution (right).

containing the sentence to be annotated. Even though reading only the sentence is sufficient for the annotation task in most cases, sometimes it is difficult to determine the correct POV without reading some preceding or following sentences. For example, if the target sentence contains a pronoun, the annotator needs the context of the paragraph to resolve the reference.

We select one word from the sentence to be annotated randomly and render it in green. In the figure, the word is *agency*. The worker has to type this word in the corresponding answer field instead of using radio buttons or check boxes. We have found that this simple copying operation improves AMT annotation quality (Laws et al., 2011).

Workers are asked to label the sentence with one of three labels: positive, neutral and negative, based on the POV of the sentence toward the target. In Figure 1, the sentence to be annotated shows a negative POV towards the target (Mel Gibson), so we would expect the worker to annotate it as negative.

Incomplete assignments where the worker submits the task without giving all the required information and suspicious assignments where the worker spends only a few seconds on the task are rejected and republished to a different worker.

We use Fleiss' κ (Fleiss, 1971) (instead of Cohen's κ) to compute intercoder agreement because it can be applied when there are more than two raters and different items are rated by different raters (which is the case when using AMT). κ is .215 for Arabic, which is considered *fair* agreement; and .419 for English, which is considered *moderate* agreement (Landis and Koch, 1977).

We assign a sentence to the class chosen by at least two of the three annotators if there is such a class. If the three annotators assign three different labels (positive, neutral and negative), we assign the sentence to the neutral class. The proportion of sentences that have agreement between two or more workers is .90 for Arabic and .95 for English. Table 1 (left) summarizes agreement statistics for the absolute POV gold standard.

The difference between the agreement among the three workers and the agreement of the majority of workers (two workers) indicates problems with the quality of the AMT results. A number of workers did not follow the instructions very carefully. For example, some workers labeled sentences that are not directly relevant to the target as positive/negative, in violation of the instructions. Our impression is that one cause of such incorrect annotations is inexperience with AMT; in general, Arabic workers seem to have less experience than English workers.

Also, as we discussed above while there are many sentences that clearly belong to a particular POV class, other sentences are in the grey area between the two. One of the authors[4] assigned labels to a subset of Arabic sentences and compared them with the labels that were assigned to the sentences based on the majority-based gold standard label. We found that gold standard labels generally agree with our own judgments.

[4] Khalid Al Khatib, a native speaker of Arabic.

Table 1 (right) shows the distribution of labels. The positive class is more frequent than the negative class in both Arabic and English. The reason seems to be that the majority of people deemed worthy of a Wikipedia article are people like inventors, poets and athletes who are generally described in a positive way.

Gold standard for POV differences. As for the gold standard for absolute POV, we have to make decisions about three aspects for the gold standard for POV differences: unit of annotation, classification scheme and type of annotator.

For POV differences, our unit of annotation is a *pair of Wikipedia articles*. We need a pair of articles because a difference can only be annotated if the two things that we want to compare are represented. We have to go up to the level of documents because Wikipedias of different languages are not aligned on the sentence/paragraph level. We use Interwiki links to establish which articles in Arabic and English correspond to each other.

We use JWPL[5] to download articles that are in the *people* category and present in both the 20120114 Arabic and the 20111115 English Wikipedia. There are 16,000 such pairs.

We selected four categories that we sampled pairs of articles from. These categories are: Arab nationalists (5 pairs), Israeli nationalists (5 pairs), hand picked (5 pairs), and random (15 pairs). The motivation for the first two categories is that we expect strong POV differences for Arab and Israeli nationalists based on our personal knowledge of the two Wikipedias. Including these ten pairs ensures that a wide spectrum of POV differences is represented in the evaluation set. For the hand picked category, we selected people who are internationally well known both in the West and the Arab world. The motivation for this category is that we want to be able to present some results to the reader that are easy to interpret – without having to look up obscure personalities in Wikipedia. The random subset (15 pairs) is a standard random sample.

The length range of downloaded articles is 1–1128 sentences for English and 1–1050 sentences for Arabic. Short articles have many problems concerning quality and completeness and are often marked as stubs that require further work. We therefore impose the constraint that both articles must contain at least 50 sentences. We also exclude very long articles because they would make the annotation task too time-consuming and expensive.

The second design decision concerns the classification scheme. Here we propose a scheme with five different classes: much more positive, more positive, equal, more negative and much more negative. This scheme is more fine-grained than for absolute POV because a document pair is a rich source of information compared to a single sentence. There is sufficient information available to make more subtle distinctions such as between “more positive” and “much more positive”.

The final design decision concerns the annotators. Here we decided against AMT because reading, understanding and evaluating a pair of documents is a complex and time-consuming task that does not correspond to the typical HIT on AMT. More importantly, we need annotators for the task that are highly proficient in both Arabic and English. This type of annotator is difficult to find on AMT; and it is difficult to verify a high level of proficiency in a language on AMT.

For these reasons, we decided to hire engineering master students at our university for the annotation task. They are all students in an information technology master’s program, native

[5] `http://www.ukp.tu-darmstadt.de/software/jwpl/`

Target: Michael Faraday

Q1. The attitude of the English article toward the target compared to the Arabic article is:

1. Much more positive
2. More positive
3. Equal
4. More negative **X**
5. Much more negative

Q2. Briefly justify your answer to question 1.

Both Articles have a very positive attitude towards Faraday but I sensed it more in the Arabic one. For example in the marriage section of the Arabic article compared with the English one ,the attitude was much more positive and it mentioned that he was a loved , devoted, humble person which isn't mentioned in the English article. Also the controversy with Davy was only mentioned in the English article not the Arabic one. The Arabic article didn't mention anything negative towards Faraday.

Figure 2: Annotation setup and example of a completed annotation for POV differences. The annotator chose "More negative" ("**X**") and wrote an explanation (in italics).

speakers of Arabic and have an excellent command of English.

The annotation setup is shown in Figure 2. The annotator reads the Arabic and English Wikipedia articles and compares the two articles based on the articles' POV toward the target. The annotation guidelines state that information that is not directly related to the target must be ignored in the annotation decision and that the decision must be based solely on the contents of the two articles. Annotators are also instructed to not be influenced by their personal opinion, emotion or POV toward the target. The annotators have to justify their answers. In our experience, this helps the annotators to provide consistent and objective annotations.

Each pair of articles is annotated by three different annotators. We map the five point rating scale to $[-2, -1, 0, 1, 2]$; e.g., "much more positive" is mapped to 2. The gold standard score Δ^{g}_{POV} for a pair of articles is then the average of the three scores given by the annotators (where the superscript g indicates "gold standard").

Intercoder agreement is $\alpha = .585$ (Krippendorff, 2004). This agreement is not as good as we would like it to be, but it is sufficient to evaluate our method; several other studies have published evaluation results based on gold standards with similar agreement (Bhardwaj et al., 2010; Brusk et al., 2010; Becker et al., 2012; Chen et al., 2012).[6]

4 Method

Absolute POV classification. For the task of determining the absolute POV – positive, neutral, negative – of a sentence, we adopt a statistical classification approach and use the Stanford MaxEnt classifier (Manning and Klein, 2003) with default parameters.

We refer to the probability of the positive (resp. negative) class for a sentence s as PosScore (resp. NegScore):

$$\text{PosScore}(s) = P(\text{positive}|s) \qquad \text{NegScore}(s) = P(\text{negative}|s)$$

Our features are bag of words (BOW) and letter k-grams (n-grams) where $2 \leq k \leq 6$.

[6] The two gold standards are available at `ifnlp.org/~schuetze/pov`.

For English, BOW and n-gram features are directly computed from text (as tokenized by the Stanford classifier) without any further linguistic preprocessing like lemmatization.

For Arabic, we investigate a number of different options for linguistic preprocessing. Arabic is a clitic language and highly inflectional. Normalization and lemmatization of Arabic text are beneficial preprocessing steps in many NLP applications. Lemmatization has been used widely in classification problems due to its ability to generate one form that matches many other related forms (Al Ameed et al., 2005). Therefore, in addition to the non-lemmatized surface forms, we used two lemmatization types: stem and root. We use *light stemming* to extract the stem: only frequent suffixes/prefixes are removed. In contrast, a word is reduced to its corresponding root by removing *all* affixes, not just frequent affixes (Al Ameed et al., 2005). We use the Arabic Text Mining tool for computing stems and roots.[7]

We use the term "bag of words" to refer to all word-level features, including "bag of stems" and "bag of roots".

Estimation of POV differences. To estimate POV differences we first need an aggregate measure of absolute POV on the document level. For this purpose, we define a document's POVScore as follows:

$\text{POVScore}(d) = 1/|d|[\sum_{s\in d}(\text{PosScore}(s) - \text{NegScore}(s))]$

where $|d|$ is the number of sentences in the document. The POVScore is simply the difference of the averages of the PosScores and NegScores of the sentences of the article. This scoring method takes into consideration how positive or negative each sentence in the article is while ignoring any neutral meaning components. The higher POVScore(d), the more positive the article is. POVScore ranges from -1 to 1.

Our assumption is that most sentences of a Wikipedia article describe the target directly. This assumption can result in errors as we will discuss in Section 6.

We can now define the POV difference Δ_{POV} of a pair of articles as the difference of the POVScore of the English article and the POVScore of the Arabic article:

$\Delta_{\text{POV}}(\{d_e, d_a\}) = \text{POVScore}(d_e) - \text{POVScore}(d_a)$

where d_e is the English article of the pair and d_a is the Arabic article of the pair.

5 Experiments and results

Absolute POV classification. We train MaxEnt in tenfold cross validation on the gold standard described in Section 3 using the BOW and letter n-gram representations described in Section 4. Folds were constructed in a way that ensures that all sentences from a particular Wikipedia article are in same fold. The baseline in our experiment is to assign all sentences to the positive class, the most frequent class in both Arabic and English.

Table 2 gives evaluation results. The best result in each column is in bold. For Arabic, the classifier is better than the baseline for all six representations, both in accuracy and F_1. The overall best results are achieved using the stem representation with letter n-grams: accuracy is .584, F_1 is .474. The problem of root-based lemmatization is that many words with the same root have different meanings (Al Ameed et al., 2005). BOW without lemmatization ("BOW, tokens") performs less well because Arabic is highly inflected.

For English, accuracy is .608 and F_1 .533 using n-grams. Using BOW, accuracy is .587 and F_1

[7] `http://sourceforge.net/projects/ar-text-mining`

		Arabic		English	
		acc	F_1	acc	F_1
	baseline	.437	.206	.478	.214
BOW	tokens	.569†	.447†	.587†	.506†
	roots	.555†	.464†		
	stems	.561†	.460†		
n-grams	tokens	.574†	.453†	**.608**†	**.533**†
	roots	.580†	.470†		
	stems	**.584**†	**.474**†		

Table 2: Accuracy (acc) and F_1 of absolute POV classification. BOW = bag of words. †: significantly better than the baseline ($p < .01$).

	POVScore		
personality	English	Arabic	Δ_{POV}
Baruch Goldstein	.244	-.199	.443
Tzipi Livni	.549	.009	.540
Ariel Sharon	.259	-.113	.372
Wael Ghonim	.398	.528	-.130
Gamal Abdel Nasser	.296	.389	-.093
Saladin	.201	.320	-.119
Michael Jackson	.475	.617	-.142
Maria Sharapova	.579	.683	-.104
Steven Spielberg	.570	.744	-.174

Table 3: POV differences: Israeli, Islamic/Arabic and international personalities.

.506. The classifier outperforms the baseline by a fair margin in this case too. The best results are achieved using n-gram features.

All differences in accuracy and F_1 between the classifier and the baseline are statistically significant at $p < .01$.[8] However, the differences in accuracy and F_1 between using BOW and n-grams features are not significant.

Estimation of POV differences. We use the classifiers that achieved the best results in the previous section for computing POV differences: n-grams on tokens for the English absolute POV classifier and n-grams on stems for the Arabic absolute POV classifier.

For each of the 30 gold standard pairs (Section 3), we run the Arabic (resp. English) classifier on all sentences of the Arabic (resp. English) article. We then compute the two document scores and the difference $\Delta_{\text{POV}}(\{d_e, d_a\})$ (Section 4).

Table 3, shows examples for three Israeli, three Islamic/Arabic and three international personalities. Israeli personalities generally have more negative articles in Arabic than English. Islamic and Arabic personalities generally have more positive articles in Arabic than in English. International personalities also have more positive articles in Arabic than in English; we attribute this to our impression that Arabic Wikipedia authors tend to be more enthusiastic about the achievements of artists and athletes even if they are held in high regard in both the Arabic-speaking and the English-speaking world.

[8] Approximate randomization test (Noreen, 1989)

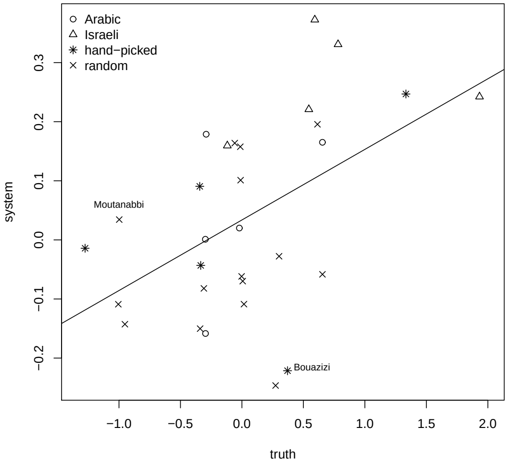


Figure 3: Estimated POV difference Δ_{POV} as a function of true POV difference Δ^{g}_{POV}. The different symbols are used to show the category of each gold standard pair.

Figure 3 plots estimated POV difference Δ_{POV} as a function of true POV difference Δ^{g}_{POV} for the 30 pairs in the evaluation set. The correlation between human annotations and system scores is statistically significant for Spearman's ρ and Kendall's τ (both at $p < .005$).

We performed an error analysis for large divergences between true and predicted POV difference. The reason for divergences mostly seems to be that we use a simple BOW representation. We will illustrate this problem with two pairs of articles – Bouazizi and Moutanabbi (see marked points in Figure 3) – in the analysis we present in the next section.

6 Discussion

In our analysis of the errors we found that most of the cases with large divergences between automatically calculated POV difference and gold standard POV difference were due to the simple BOW representation we use – where we will use *BOW* in this section as a short hand for both bag of words and n-gram representations. This type of classifier is often not capable of detecting the subtle semantic nuances that are necessary to accurately assess absolute POV and POV differences. There are two main subcases of this general problem.

First, our assumption that all sentences in the Wikipedia article are describing the target is incorrect. (As in the rest of the paper we refer to the subject of the Wikipedia article as the target in this section.) There are sentences that describe people or events that have an impact on the life of the target, but are not directly about the target. These sentences can affect the system POVScore even though they are not communicating information relevant to the absolute

POV of the article towards the target. Identifying the subject of a sentence (target vs. something else) is not possible using BOW.

A special case of this are passages in Wikipedia about artists that contain titles and descriptions of movies, novels and other works of art. Again, this information can affect the system POVScore even though the fact that – for example – an actor played a murderer does not contribute to a negative POV about him.

The second subcase concerns parts of articles that *are* directly about the target, but not relevant for POV. The article may describe negative events that happened to the target, e.g., "Roosevelt contracted ... polio which resulted in permanent paralysis." Again, this will decrease the POVScore of the article even though the information reports something negative about the circumstances of the person's life that will not affect a reader's POV towards the person in a negative way.

An even subtler problem occurs if positive or negative words occur in a sentence that is directly relevant for POV towards the target, but these positive or negative words are in the scope of another word that reverses their meaning (cf. (Kessler and Schütze, 2012)). For example, the statement: *John started a war on violence against women* supports a positive POV towards John even though most of the words in the statement are negative words.

The immediate effect of the shortcomings of a BOW-based feature representation is an incorrect estimation of absolute POV. However, since these effects are somewhat random and will in most cases not affect Arabic and English to the same extent, the BOW problem can also give rise to incorrect POV differences.

In our data set, this is the reason that our system does not correctly predict the POV difference for Mohamed Bouazizi, the Tunisian who is credited with starting the Arab Spring (see data point marked "Bouazizi" in Figure 3). The system prediction is -.217 whereas the true score is .333. The problem with this pair of articles is that Bouazizi is described as a mostly positive person in both languages, but the circumstances of his life are described as tragic. Since our system does not distinguish between sentences that are directly relevant about the target vs. those that are not, this causes an incorrectly estimated POV difference.

A second example is Moutanabbi, a famous Arabic poet (system score: .034, truth: -1.00, data point marked "Moutanabbi" in Figure 3). His English article is positive, but his Arabic article is even more positive, hence the truth score of -1.00. The Arabic article is not handled well by a BOW representation for similar reasons as for Bouazizi. In particular, it contains poems about negative phenomena like mudslinging and sadness; and it describes the negative behavior of fellow poets towards Moutanabbi – this is negative, but will not create a negative impression of Moutanabbi in the mind of the reader.

There is one positive aspect of simplistic BOW representations. A potential concern is that the annotation of POV could be affected by annotator bias. Annotators have their own POV and even though we explicitly ask them to base their annotations solely on the content provided, there is a danger that they will be influenced by their personal views. However, in a BOW model this is not a problem: potentially incorrect annotations may contribute noise, but no systematic biases will be introduced. For example, even if an annotator is sympathetic with a murderer and the resulting annotation could mislead a classifier into believing that "murderer" is a positive word, there will be other annotations containing "murderer" that will counterbalance the incorrect annotation.

Note that for POV *differences*, personal annotator POV is less of a problem. We can expect a good annotator to provide a high-quality assessement of POV differences because a relative judgment about two articles is not in conflict with one's own personal views.

7 Related Work

Sentiment analysis (Pang and Lee, 2008) is mostly concerned with subjective language. However, the classification of objective language into positive vs. negative might also be considered sentiment analysis since the two tasks have a similar structure and face similar challenges.

Many papers have studied sentiment analysis in the news domain. Although this domain mainly contains objective content, subjectivity is also found to some extent, e.g., in editorials. Most previous work on sentiment analysis of news has ignored objective content. For example, Wiebe et al. (2005) released the Multi-Perspective Question Answering (MPQA) corpus. This corpus has detailed manual annotations of a set of 535 news articles. The corpus separates subjective and objective expressions. It has some information about objective content (such as the source and the target of the objective speech), but only has sentiment information about the subjective content. Our annotated corpus is different because it is concerned with objective language. Our task requires different annotation guidelines and, in general, a different setup for the annotation process compared to work on sentiment analysis.

Balahur et al. (2010), Abdul-Mageed and Diab (2011) and Balahur and Steinberger (2009) try to distinguish between positive and negative sentiment vs. good and bad news. Good and bad news are considered objective information and excluded from the classification process. In contrast, our method deals with good and bad objective information in the classification step.

Some prior work has classified financial news according to polarity. Some papers limit their classification to the subjective content of the news (e.g., (Agic et al., 2010)); other papers have classified objective content as well (Ahmad, 2006; Devitt and Ahmad, 2007; Shtrimberg, 2004). For example, Shtrimberg (2004) proposed an approach to classify news stories about companies as positive or negative. His classifier learned from a corpus where every news story about a company is labeled based on its impact on the future price of its stock. However, impact on price is different from positive/negative. For example, bad economic news can have a positive impact on stock prices if investors think it will make the Federal Reserve more likely to launch another round of quantitative easing. Our approach generates a score that indicates the POV of the article toward the subject matter and that is not directly related to the impact such information might have on the financial markets.

Another topic related to POV is media bias (Gentzkow and Shapiro, 2005, 2006). Some studies on this topic investigate bias in Wikipedia. Herzig et al. (2011) propose a novel annotation scheme as a basic step towards an automatic machine learning system to detect biased language in English Wikipedia. The scheme has multiple levels of bias tagging: the intra-sentential level, which includes polar-phrase, weasel, repetition, and personal-tone, and the sentence and entry level. The proposed scheme was applied to a set of articles from the service providers category in Wikipedia. Annotation categories distinguished between biased language and unbiased language. The authors conducted their annotation scheme based on the articles which explicitly *violate* the NPOV principle. Our approach studies POV differences under the assumption that Wikipedia articles mostly *adhere* to the NPOV principle and do not use biased linguistic expressions.

A line of research related to POV is work on perspectives and viewpoints (Lin et al., 2006; Paul et al., 2010). Most of this work uses Bitterlemons, a corpus of 594 articles each of which is written either from an Israeli or from a Palestinian perspective. Perspective classification (Lin and Hauptmann, 2006; Greene and Resnik, 2009; Klebanov et al., 2010) and modeling (Ahmed and Xing, 2010; Hardisty et al., 2010) then attempts to automatically detect the perspective of an article. Perspective and positive/negative POV are related, but different concepts; e.g., the sentences "political prisoners are released in Hamas deal" and "parties discuss new construction in Judea and Samaria" are both neutral or positive, but indicate different – Palestinian vs. Israeli – perspectives. In addition, much of the content of the Bitterlemons corpus is subjective – 66% of sentences according to Lin et al. (2006). In contrast, we address the problem of identifying positive/negative POV in objective language. Finally, the computational work on Bitterlemons is mostly on the document level whereas the measures we propose are based on sentences.

Massa and Scrinzi (2011) describe Manypedia, a web tool that supports comparing articles on the same subject in Wikipedia versions of different languages. They define the linguistic point of view as the potential difference in POV between Wikipedia articles from different languages due to the isolation of editor communities of these language versions. The tool provides users with multiple options such as translating the articles using Google Translate, extracting the most frequent words and showing information about editing of the article (e.g., total number of edits and editors). In contrast to our approach, Manypedia does not aim to provide automatic NLP analysis functionalities.

8 Conclusion and Future Work

The comparative analysis of differences – be they subtle or conspicuous – in the evaluation of particular subject matters is of great importance in the social sciences. The method we propose in this paper has two advantages. By choosing a relative instead of an absolute approach we avoid the old and still unsettled problem of defining an objective, neutral standard; and by taking a statistical classification approach, we provide an automatic method suitable for the analysis of large amounts of text.

Future work. We would like to address two problems in future work. First, our error analysis showed that most errors in predicting POV difference were due to our simple representation of sentences: bag of words. We would like to use more sophisticated representations that take into account the scope of positive and negative words; and also language understanding methods that can detect what a statement is about – the target itself or something not directly related to the target.

Second, we pointed out that there seem to be cultural differences in the magnitude of absolute POV. In particular, we found that international personalities are viewed in more positive light in the Arabic Wikipedia than in English. This means that there are at least two possible reasons for a POV difference: it can be due to a generally lower or higher level of absolute POV in one language; or it can be due to a genuinely different evaluation of a personality in two Wikipedias. We plan to distinguish these two different kinds of POV difference in future work.

Acknowledgments. We thank Deutsche Forschungsgemeinschaft for funding this research (DFG, SFB 732, D7); the anonymous reviewers, Christian Scheible, Wiltrud Kessler, Charles Jochim and Andrea Glaser for their valuable comments; and Amr Yassin, Ghada Dessouky, Mariam Hassib and Mirna Bouchra for performing the annotation.

References

Abdul-Mageed, M. and Diab, M. T. (2011). Subjectivity and sentiment annotation of modern standard Arabic newswire. In *Linguistic Annotation Workshop*, pages 110–118.

Agic, Z., Ljubesic, N., and Tadic, M. (2010). Towards sentiment analysis of financial texts in croatian. In *The 7th International Conference on Language Resources and Evaluation*, LREC '10.

Ahmad, K. (2006). Multi-lingual sentiment analysis of financial news streams. *PoS*, GRID '06:001.

Ahmed, A. and Xing, E. P. (2010). Staying informed: Supervised and semi-supervised multi-view topical analysis of ideological perspective. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1140–1150.

Al Ameed, H. K., Al Ketbi, S. O., Al Kaabi, A. A., Al Shebli, K. S., Al Shamsi, N. F., Al Nuaimi, N. H., and Al Muhairi, S. S. (2005). Arabic light stemmer: A new enhanced approach. In *Proceedings of the 2nd International Conference on Innovations in Information Technology*, IIT '05.

Alonso, O. and Lease, M. (2011). Crowdsourcing for information retrieval: Principles, methods, and applications. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 1299–1300.

Balahur, A. and Steinberger, R. (2009). Rethinking sentiment analysis in the news: From theory to practice and back. In *Proceedings of the 1st Workshop On Opinion Mining and Sentiment Analysis*.

Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., van der Goot, E., Halkia, M., Pouliquen, B., and Belyaeva, J. (2010). Sentiment analysis in the news. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, LREC '10.

Becker, L., Basu, S., and Vanderwende, L. (2012). Mind the gap: Learning to choose gaps for question generation. In *The 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '12, pages 742–751.

Berger, P. L. and Luckmann, T. (1966). *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. Anchor books. Doubleday.

Bhardwaj, V., Passonneau, R. J., Salleb-Aouissi, A., and Ide, N. (2010). Anveshan: A framework for analysis of multiple annotators' labeling behavior. In *Proceedings of the 4th Linguistic Annotation Workshop*, LAW IV '10, pages 47–55.

Brusk, J., Artstein, R., and Traum, D. R. (2010). Don't tell anyone! Two experiments on gossip conversations. In *Proceedings of The 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '10, pages 193–200.

Chen, J., Ding, R., Jiang, S., and Knudson, R. (2012). A preliminary evaluation of metadata records machine translation. *The Electronic Library*, 30(2).

D'Alessio, D. and Allen, M. (2000). Media bias in presidential elections: A meta-analysis. *Journal of Communication*, 50:133–156.

Devitt, A. and Ahmad, K. (2007). Sentiment polarity identification in financial news: A cohesion-based approach. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, ACL '07.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Früh, W. (2011). *Inhaltsanalyse. Theorie und Praxis*. UTB.

Gentzkow, M. and Shapiro, J. (2005). Media bias and reputation. Working Paper 11664, National Bureau of Economic Research.

Gentzkow, M. and Shapiro, J. M. (2006). What drives media slant? Evidence from U.S. daily newspapers. Working Paper 12707, National Bureau of Economic Research.

Greene, S. and Resnik, P. (2009). More than words: Syntactic packaging and implicit sentiment. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 503–511.

Habermas, J. (1984). *The Theory of Communicative Action. Reason and the Rationalization of Society*, volume 1. Heinemann.

Habermas, J. (2006). Political communication in media society: Does democracy still enjoy an epistemic dimension? The impact of normative theory on empirical research. *Communication Theory*, 16:411–426.

Hardisty, E. A., Boyd-Graber, J., and Resnik, P. (2010). Modeling perspective using adaptor grammars. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 284–292.

Herzig, L., Nunes, A., and Snir, B. (2011). An annotation scheme for automated bias detection in Wikipedia. In *Proceedings of the 5th Linguistic Annotation Workshop*, LAW V '11, pages 47–55.

Kessler, W. and Schütze, H. (2012). Classification of inconsistent sentiment words using syntactic constructions. In *Proceedings of the 24th International Conference on Computational Linguistics*, COLING 24.

Klebanov, B. B., Beigman, E., and Diermeier, D. (2010). Vocabulary choice as an indicator of perspective. In *Proceedings of the The 48th Annual Meeting of the Association for Computational Linguistics*, pages 253–257.

Krippendorff, K. (2004). *Content Analysis: An Introduction to Its Methodology*. Sage Publications.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Laws, F., Scheible, C., and Schütze, H. (2011). Active learning with Amazon Mechanical Turk. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1546–1556.

Lin, W.-H. and Hauptmann, A. (2006). Are these documents written from different perspectives? A test of different perspectives based on statistical distribution divergence. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL '06, pages 1057–1064.

Lin, W.-H., Wilson, T., Wiebe, J., and Hauptmann, A. (2006). Which side are you on? Identifying perspectives at the document and sentence levels. In *Proceedings of the 10th Conference on Computational Natural Language Learning*, CoNLL-X '06, pages 109–116.

Littlejohn, S. W. and Foss, K. A. (2010). *Theories of Human Communication*. Waveland.

Manning, C. and Klein, D. (2003). Optimization, maxent models, and conditional estimation without magic. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Tutorials - Volume 5*, NAACL-Tutorials '03.

Massa, P. and Scrinzi, F. (2011). Exploring linguistic points of view of Wikipedia. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, WikiSym '11, pages 213–214.

Noreen, E. (1989). *Computer-Intensive Methods for Testing Hypotheses: An Introduction*. Wiley.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135.

Paul, M. J., Zhai, C., and Girju, R. (2010). Summarizing contrastive viewpoints in opinionated text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 66–76.

Scheufele, D. A. (1999). Framing as a theory of media effects. *Journal of Communication*, 49(1):103–122.

Shtrimberg, I. (2004). Good news or bad news? Let the market decide. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text*, pages 86–88.

Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210.