An Empirical Study for Determining Relevant Features for Sentiment Summarization of Online Conversational Documents

Gino Mangnoesing[1], Arthur van Bunningen[2],
Alexander Hogenboom[1], Frederik Hogenboom[1], and Flavius Frasincar[1]

[1] Erasmus University Rotterdam
PO Box 1738, NL-3000 DR, Rotterdam, the Netherlands
gvh.sing@gmail.com, {hogenboom, fhogenboom, frasincar}@ese.eur.nl
[2] Teezir BV
Wilhelminapark 46, NL-3581 NL, Utrecht, the Netherlands
Arthur.van.Bunningen@teezir.com

Abstract. The phenomenon of big data makes managing, processing, and extracting valuable information from the Web an increasingly challenging task. As such, the abundance of user-generated content with opinions about products or brands requires appropriate tools in order to be able to capture consumer sentiment. Such tools can be used to aggregate content by means of sentiment summarization techniques, extracting text segments that reflect the overall sentiment of a text in a compressed form. We explore what features distinguish relevant from irrelevant text segments in terms of the extent to which they reflect the overall sentiment of conversational documents. In our empirical study on a collection of Dutch conversational documents, we find that text segments with opinions, segments with arguments supporting these opinions, segments discussing aspects of the subject of a text, and relatively long sentences are key indicators for text segments that summarize the sentiment conveyed by a text as a whole.

1 Introduction

In the last decade, the World Wide Web has exponentially grown to a network with more than 555 million websites and over 2 billion users worldwide [11]. The Web has become an influential source of information with an increasing share of user-generated content (UGC) from many contributors. This content has taken many different forms, such as forums, wikis, (micro)blogs, review sites, podcasts, or parts of a website, e.g., reviews on Amazon.com or Booking.com. Amongst all this content are many opinions which can carry a specific sentiment, for example about products, brands, or politics. People complain or recommend what products or services to buy or not to buy, which movies to see, or what places to go to. Consequently, the Web as a medium has become a strong influencer of purchasing decisions and a platform that reflects consumer preferences, which is interesting for both consumers and producers.

However, it has become difficult to use the Web as a helping hand for making decisions, as it has become much harder to keep track of all available data online. Today's data is often unstructured, scattered all over the Web, and expanding extremely fast. This phenomenon is also referred to as big data [7]. There is simply too much information to process, as well as a lack of filters to extract the parts that are relevant and informative with respect to one's requirements.

This situation has led to a great need for aggregation for a better information overview and making big data insightful and eventually profitable. Fortunately, there are ways to accomplish aggregation of opinions by means of sentiment summarization. Whereas sentiment analysis computes a score to indicate the attitude people have towards a certain topic, sentiment summarization takes it one step further by extracting the most important text fragments that sufficiently represent the sentiment of the text as a whole. In this light, sentiment summarization could help consumers to quickly discover the pros and cons of products and services and it could support companies with brand monitoring and customer relationship management.

Existing work in the field of sentiment summarization identifies relevant text fragments by using one or just a few characteristics (features), such as product aspects [3, 6, 12–14], intensity [6], or a fragment's position within a document [2]. In our current work, we take a much broader approach, attempting to learn about the relative importance of a larger set of features. Additionally, rather than focusing on automatic detection of features, we let users annotate features in order to find out which features apply to a specific summary. Another distinctive factor in our study is our focus on conversational documents about a brand, taken from forums, whereas many existing studies use opinion-focused documents like movie reviews or restaurant reviews. Conversational documents are texts that have the characteristics of a dialogue, for example texts in which people ask or answer questions, give comments, or express complaints. Conversational documents are typically found on forums and social media platforms.

The remainder of this paper is structured as follows. Section 2 discusses related work on sentiment summarization. Section 3 discusses the features we consider as proxies for the relevance of sentences in terms of the extent to which they reflect the overall sentiment of a text. Section 4 discusses our method for feature evaluation of sentiment summaries. The evaluation of the proposed method is discussed in Section 5. Finally, we present our conclusions and directions for future work in Section 6.

2 Sentiment Summarization

Sentiment analysis typically aims to examine "what other people think" about a specific entity or topic [9, 10] and allows for determining a score for the polarity of text. These sentiment scores are important for sentiment summarization, as the aim here is to extract text fragments that reflect the overall sentiment of a text, while using less words than the original text. The key to generating such a summary is distinguishing relevant sentences from irrelevant sentences.

Existing work is typically focused on sentences that discuss aspects of a topic and often summarize sentiment found in reviews. For instance, Blair-Goldensohn et al. [3] present a sentiment summarizer for user reviews of local services like restaurants or hotels. Their method extracts the snippets from the text that include sentiment-carrying text. The snippets are then examined on the presence of service aspects, such as the food, service, or price. Subsequently, the sentiment per aspect is aggregated, based on the snippets that include these aspects. Blair-Goldensohn et al. [3] observe that most services share similar basic aspects and that acceptable summaries can be created when guiding the selection process of text fragments by the presence of topic aspects in these fragments. However, other work has shown that other features have their merit too [2, 6].

Lerman et al. [6] propose several sentiment summarizers that include various summary features. One of the used features besides aspects is the intensity of a text fragment, capturing the magnitude of its conveyed sentiment. Another considered feature is the mismatch of a text fragment, which measures the difference between the sentiment of the fragment and the known overall sentiment of the topic. The results of this study indicate that none of the proposed sentiment summarizers is strongly preferred over any other. However, users prefer summarizers that account for aspects and sentiment over those that do not.

Beneike et al. [2] aim to extract a single fragment from a movie review that reflects the sentiment of the author towards the movie. This fragment is referred to as a quotation. Beneike et al. [2] show that three features appear to be predictive of whether a text fragment is chosen as a quotation. The first feature is the location of the fragment within the paragraph. Quotations occur most often at the ends of paragraphs. The second feature is the location of the fragment within a document, often early in the document or the final 5% of the text. The third feature is the word choice. In quotations, the most used words often express emotion directly and/or are interchangeable with the topic, e.g., a reviewer with a positive opinion on a movie may refer to the movie as "a piece of art".

As such, existing work only considers few features as proxies for the relevance of a text fragment for a sentiment summary. Existing work shows only few common denominators in considered features, with one of the most widely used features being the discussion of certain aspects of a topic. We aim to investigate which of the existing as well as new features can identify relevant fragments for sentiment summaries. In addition, the focus of our work is on conversational documents captured by forum messages, which differ from the reviews typically used in existing work in that they are less explicitly focused on expressing opinions and in that they are structured as conversations rather than as single messages.

3 Selecting Features for Sentiment Summarization

In this research, we perform a user-based evaluation of summary sentences and various summarization features. Users evaluate these sentences in terms of the extent to which they would fit in a summary reflecting the sentiment conveyed by the text as a whole. The features we explore in this research are listed below.

1. The sentence contains an opinion about the topic.
2. The sentence is rather positive or rather negative (high intensity).
3. The sentence includes one or more (sub)aspects of the topic.
4. The sentence is part of the introduction of the document.
5. The sentence is part of the conclusion of the document.
6. The sentence contains an adjective.
7. The sentence contains an adverb.
8. The sentence addresses an event or experience described in the document.
9. The sentence contains an advice or recommendation.
10. The sentence contains an argument supporting an opinion, vision, or statement in the document.
11. The sentence contains or is part of a comparison in the document.
12. The sentence contains words that are also present in the document title (with the exception of definite and indefinite articles).
13. The sentence contains a list or sequence.
14. The sentence is relatively long.
15. The sentence is relatively short.

Some of these features are inspired by existing work, discussed in Section 2, in order to be able to validate existing results on conversational documents. Yet, most of our considered features are contributions of our current endeavors. We include feature 8, as we hypothesize that discussions of events are relevant on forums, because people tend to use forums to complain about something, which is typically a recent experience or event. Feature 9 is considered, because we hypothesize that a recommendation can be important for a summary. For example, an advice may reflect or be supported by someone's sentiment towards the brand. Similarly, arguments can further motivate an opinion or statement, thus rendering feature 10 an interesting feature. In this light, we also consider feature 11, as people tend to motivate why they favor one brand over another in a comparison. Additionally, feature 12 is included, because the subject of forum conversations is typically mentioned in the title of a document. Titles may hence form a concise representation of the most important message of a document and may thus be useful in identifying relevant fragments in the body of a document. Feature 13 is included as people tend to list their complaints and possible advantages or disadvantages. Last, features 14 and 15 are included in order to determine whether relevant summary sentences are typically relatively long or short sentences.

4 Feature Evaluation for Sentiment Summarization

In order to support our current research goals, we propose a method for Feature Evaluation for Sentiment Summarization (FESS). The goal of this method is to collect user evaluations of sentences with respect to the relevance of the sentence for inclusion in a sentiment summary. By doing so, we aim to find out which features are good indicators for the relevance of a sentence. In the following sections, we first present an overview of our method design. Then, we elaborate on the implementation of our method.

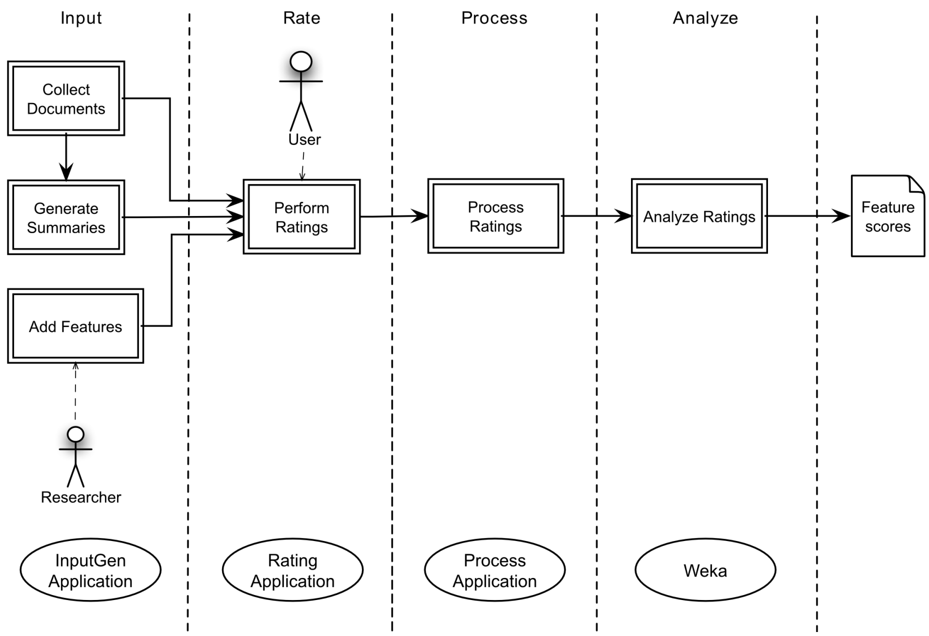

Fig. 1. A schematic overview of our proposed method.

4.1 Method Design

Figure 1 demonstrates the design of our proposed method. FESS consists of four main steps. First, we prepare the inputs for the second step, in which we collect evaluations for these inputs from human raters. We then process these ratings and analyze the results in order to be able to provide recommendations on which types of features to use in sentiment summarizers for conversational documents.

Input The first step of our method is intended to provide input to the second step in which we retrieve user evaluations. The first input we collect is a set of documents, crawled from the Web and filtered on several specific characteristics, e.g., the topic and the source of the document. Our method supports several types of conversational texts, e.g., tweets, comments, or forum posts.

The second input consists of candidate summary sentences. A collection of such sentences is generated for every single document in our collection. In order to do so, we first preprocess the text by detecting sentences and words. After preprocessing, we select summary candidates by flagging (marking) sentences that are likely to contain one or more of our considered summary features. Sentences may carry multiple flags. For every document, we then select a predefined number of summary sentences that are randomly selected from a collection of candidate sentences. Flags may be taken into account in this process in order to ensure that no features are under-represented in the final selection.

Last, a list of considered features of sentences is required as an input for the rating phase. We propose to use the features detailed in Section 3. When rating the selected sentences with respect to the extent to which they are relevant for inclusion in a summary, users can annotate these sentences for the presence or absence of our considered features.

Rate In the rating process, all documents, summary sentences, and features are presented to human raters. We propose to divide users into groups in order to spread the time-consuming evaluation work among users. Each group of raters is presented a unique set of documents, with several summary sentence candidates per document. The evaluation scores can be saved for further processing.

The interaction model of FESS for collecting feature evaluations is composed of four steps. First, users are presented a random document related to a predefined topic, e.g., a document relevant for a search query for a brand. After having read the document, users are presented a potential summary sentence extracted from the same document. Subsequently, we ask users to classify the sentiment conveyed by the document as well as the sentiment conveyed by the sentence. Here, one can consider classes like negative (-1), neutral (0), and positive (1). Additionally, we ask users to evaluate the relevance of this sentence in terms of the extent to which it summarizes the sentiment conveyed by the document as a whole. This can be done by means of binary classification or by means of classification on an ordinal or continuous scale that can be mapped to binary classes. Subsequently, the users are asked to select the features applying to this specific sentence from a list.

Process In the processing phase of our method, the evaluation data is transformed into a collection of data points, where each data point represents a single summary sentence with the evaluation scores for the document sentiment, the summary sentiment, the summary relevance score, and a binary representation of features, indicating whether the users selected these features (1) or not (0). In order to generate these data points, the user ratings for each sentence are aggregated. Sentiment evaluations are averaged over all raters, whereas majority voting is applied to both the (binary) relevance scores and the features.

Analyze The collection of data points thus obtained can subsequently be analyzed in order to identify important proxies for the relevance of a sentence in a sentiment summary. First, we propose to use the collected data in order to determine the information gain [8] of each considered feature with respect to the binary classification of the relevance of our evaluated sentences. As an alternative method, we propose to use a feature selection method in order to identify the most informative subset of features, by considering the predictive power of each individual feature along with the degree of redundancy between the features [5]. Subsets of features that are highly correlated with the class while having low inter-correlation are preferred. In order to validate the results, we propose to perform stratified 10-fold cross-validation on these analyses.

4.2 Implementation

We have implemented our method using programming languages ASP.NET and C# in combination with an SQL database. As depicted in Fig. 1, each step of our framework is performed by a separate application.

Input In order to provide inputs for the rating process, we have developed the InputGen Application, which collects documents, selects candidate summary sentences, and provides a list of considered features. In the document collection step, our implementation focuses on crawling popular Dutch forums for messages about Ziggo, a national media and communications services provider in the Netherlands. We filter the data for documents with a maximum length of a predefined number of characters. All documents are saved into a database.

Sentences that are potential candidates for inclusion in a sentiment summary are selected by the InputGen Application by first tokenizing the text into separate sentences and words. Then, we create a sub-collection of candidate sentences for each document and automatically flag the sentences, when they include a feature from a subset of features we scan for in advance. The collection of candidate sentences is retrieved by automatically filtering sentences with a predefined minimum amount of words. We make use of this minimum in order to filter out the majority of presumably meaningless short sentences.

Additionally, we automatically flag sentences that match the criteria of some of our features, such that features are as well-represented as possible in our selection of candidate sentences. The first feature we explicitly look for in sentences is the presence of aspects of our topic. In our application, sentences that contain words matching a pre-compiled list of lexical representations of aspects of our topic are flagged for this feature.

Another feature we explicitly look for is whether sentences are part of a conclusion. In order to accomplish this, we flag sentences in the last 25% of a document as being part of a conclusion. By doing so, we assume that conclusions typically occur at the end of the relatively short documents in our collection.

A third feature our application automatically scans for is sentences with high intensity, as high intensity often signals a strong presence of opinion. To this end, our application scores the sentences in our documents for their conveyed sentiment and assumes high absolute sentiment scores to signal high intensity. The sentences in our data set are analyzed for the sentiment conveyed by their text by means of an existing framework for lexicon-based sentiment analysis [1], which is a pipeline in which each component fulfills a specific task in analyzing the sentiment of a document. It first prepares documents by cleaning the text and performing initial linguistic analysis by identifying each word's part-of-speech as well as by distinguishing opinionated words and their modifiers from neutral words. Sentences are subsequently scored by sum-aggregating the sentiment scores of its opinionated words, while accounting for their modifiers, if any.

Except for collecting documents and automatically selecting candidate summary sentences, our InputGen Application also enables one to specify a list of considered features. These features can be manually added through an interface.

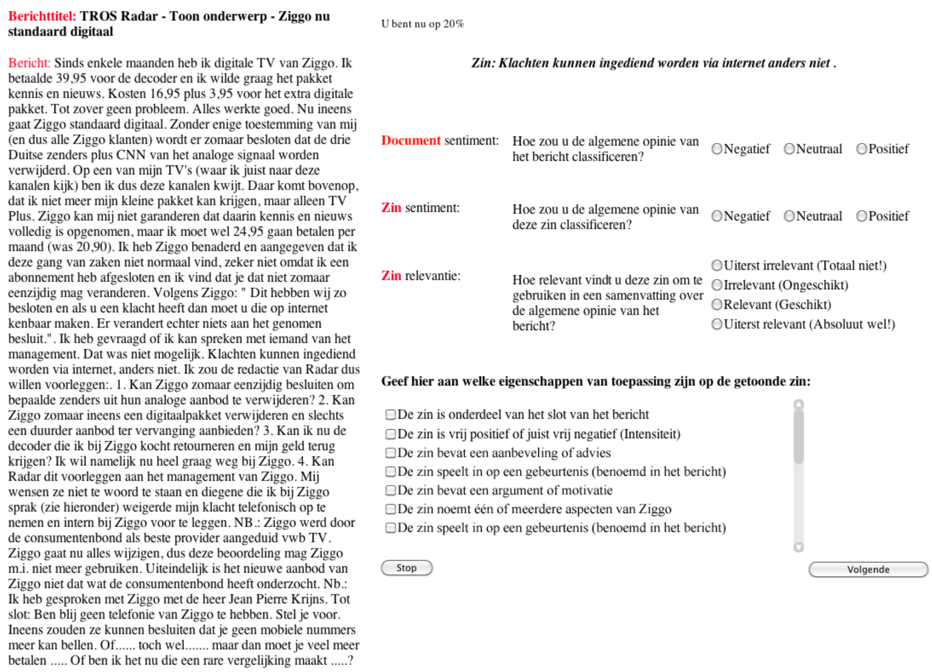

Fig. 2. Rating Application user interface.

Rate After providing all the necessary input, we let users evaluate the selected summary sentences. First, users are assigned a group number, which they can fill in at the start screen. The set of documents and summary sentences presented to a user depends on the group number. Users provide their ratings through the Rating Application, depicted in Fig. 2.

On the left hand side, we display a document. On the right hand side, we ask the user to rate the given document and a selected sentence (displayed in the upper right corner) for sentiment and relevance with respect to sentiment summarizations, as well as to select the features that apply to the selected sentence. In our current endeavors, we only focus on the relevance score and the features – sentiment scores, which can be either negative, neutral, or positive, are collected for future research purposes. The relevance of candidate summary sentences can be scored as either very irrelevant, irrelevant, relevant, or very relevant, yet these scores are mapped to relevant and irrelevant in the processing phase.

Process In order to transform the evaluation data from the rating phase into usable data, our Process Application first retrieves all evaluations from our database. Then, we generate data points by applying the majority rule, as detailed in Section 4.1. The data thus generated is saved in a format that allows for easy analysis in our application.

Analyze The analysis of our data is performed by means of the Weka software package [4]. We compute the information gain using the *InfoGainAttributeEval* method and we select subsets of the most relevant features using the *CfsSubsetEval* method combined with the *ExhaustiveSearch* method. Both analyses are performed with stratified 10-fold cross-validation.

5 Evaluation

Following the method described in Section 4, we have performed an empirical study in order to determine relevant features for sentiment summarization of online conversational documents. The experimental setup and results of this study are detailed below.

5.1 Experimental Setup

In our current endeavors, we focus on a set of 60 Dutch forum posts about the Dutch company Ziggo. We limit the size of the documents in our set to 2,500 characters. The minimum length of a candidate summary sentence is assumed to be five words. For each of our 60 documents, we select candidate summary sentences to be presented to users.

Ideally, we would present all sentences to a user. However, this would render the human evaluation phase a very time-consuming process. Therefore, we present each user only seven sentences per document, i.e., two sentences flagged for the aspects feature, one sentence flagged for intensity, one sentence flagged for being part of a conclusion, and three random sentences. Through this preselection process, we aim to reduce the risk of features being under-represented in the final selection.

For our evaluations, we divide our collection of 60 Dutch conversational documents and their associated candidate summary sentences (seven per document) into three equally-sized groups. Each group of documents is rated by a group of three human annotators. As such, we have nine human raters that evaluate 20 documents each, with seven summary sentences per document. This yields a total of 1,260 ratings, which are represented by 420 data points. These data points represent the evaluations of our human annotators for each of our considered sentences, as determined by means of majority voting.

5.2 Experimental Results

By using our application, we have collected user evaluations of seven candidate summary sentences for each of our 60 documents. The distribution of the features over our data set, according to our human annotators, is visualized in Fig. 3. Most features appear to be sufficiently represented in our data. The comparison and sequence features do however appear to be very rare in our data.

When we analyze the features in our data set in terms of their associated information gain, we can clearly distinguish useful from less useful features.

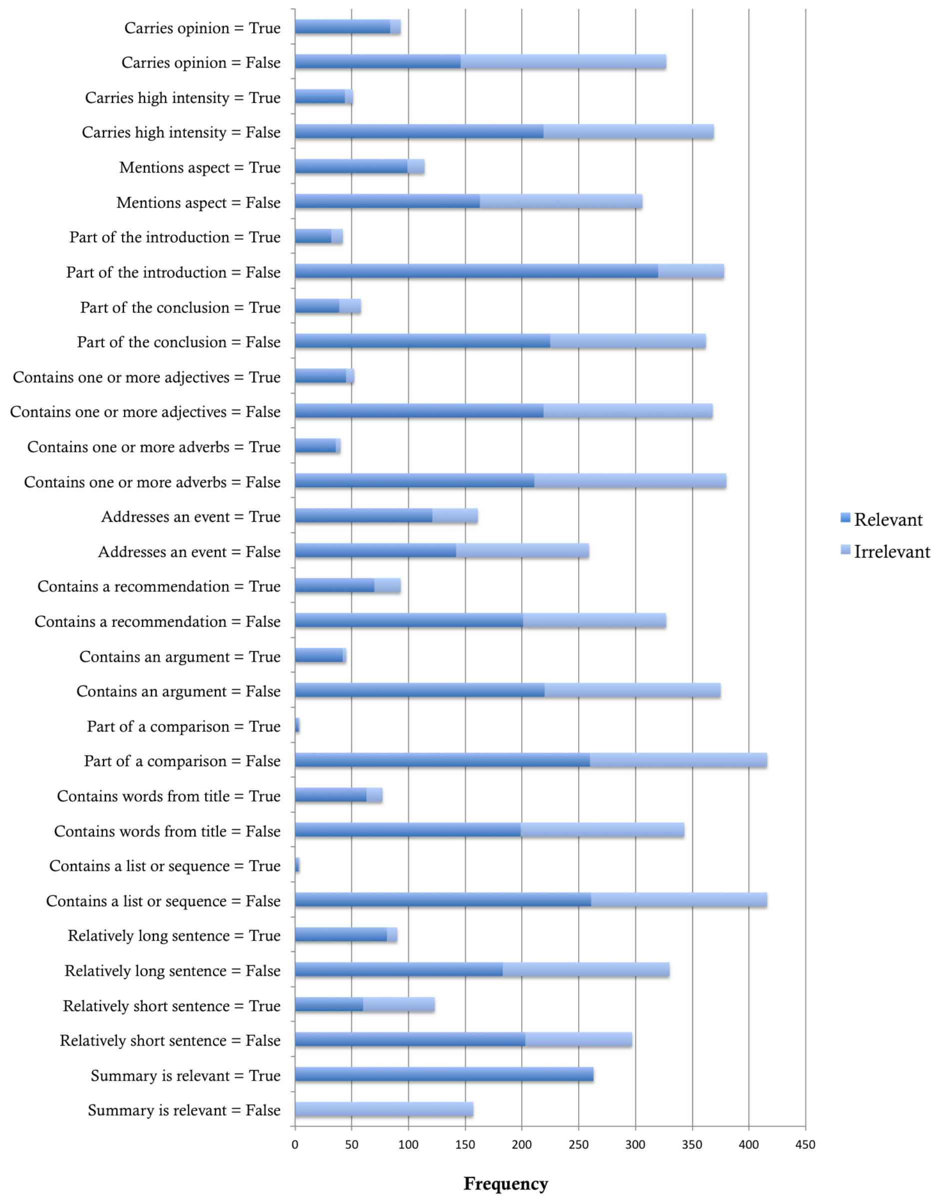

Fig. 3. Distribution of features over our data set.

Figure 4 suggests that in our data, three features contain relatively much information that can be used to distinguish relevant from irrelevant text fragments for sentiment summaries. Aspects provide the highest information gain, closely followed by sentences that are relatively long and sentences that carry opinion.

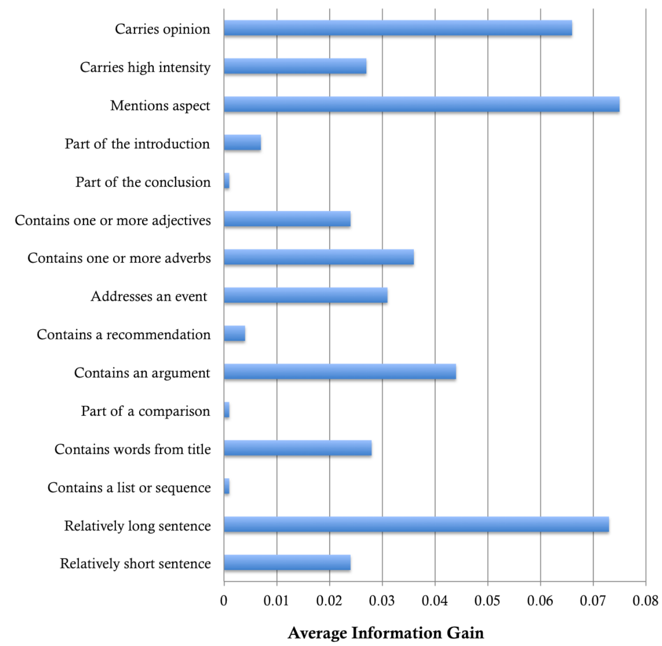

Fig. 4. Information gain of our considered features, averaged over 10 folds.

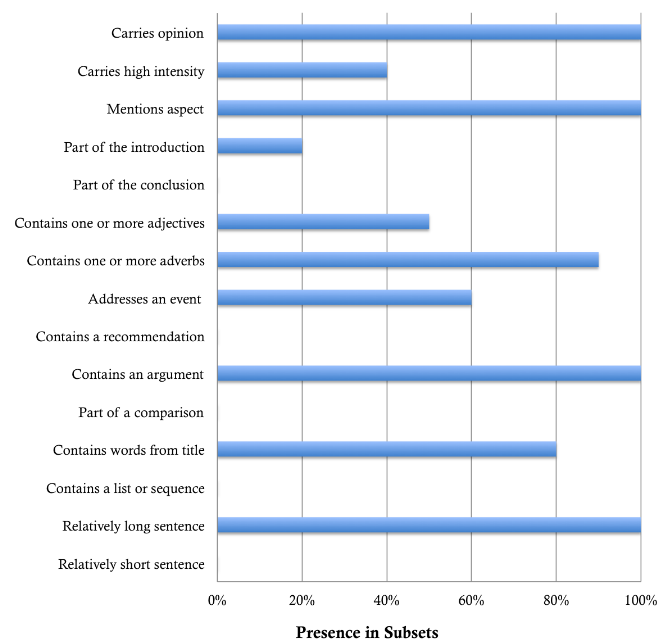

Fig. 5. Relative frequency of selection of our considered features in 10 folds.

Conversely, three features seem to have a rather marginal relevance, i.e., conclusion sentences, sentences with lists, and sentences that are part of a comparison. This may be related to their relatively low frequency in our data set.

Another analysis, in which we apply a feature selection method in order to identify the most informative subset of features in each of our 10 folds by considering the predictive power of each individual feature along with the degree of redundancy between the features [5], we obtain similar results. Figure 5 shows that four features are always selected, i.e., sentences with opinions, sentences with aspects, sentences with arguments, and long sentences. Sentences containing adverbs and sentences containing words that also occur in the title of a document are selected relatively often, yet not in all cases.

As such, our results indicate that, according to our human annotators, four features are relatively important proxies for a text fragment's relevance in sentiment summaries of conversational documents, i.e., fragments discussing aspects of a text's subject, fragments that are relatively long, fragments with opinions, and fragments containing arguments supporting these opinions. Especially features referring to arguments and long text fragments are remarkable, as they are, to the best of our knowledge, not used in existing sentiment summarizers.

6 Conclusions and Future Work

When it comes to summarizing the sentiment conveyed by a piece of conversational text, we have shown that relatively long, opinionated fragments are good candidates for inclusion in a summary. Additionally, our results indicate that in our corpus of Dutch conversational documents, it is not so much the absolute position of text fragments – e.g., fragments' occurrence in an introduction or conclusion – that distinguishes relevant from irrelevant fragments. Conversely, it is rather the role sentiment-carrying fragments play – e.g., arguments supporting the overall message, or fragments discussing different aspects of the topic – that renders them useful in summaries reflecting the sentiment of a document.

In this light, we plan to validate our findings on other corpora and languages, as well as to further investigate how we can account for structural features (e.g., argumentation structures) and semantic features (e.g., distinct aspects of a topic) of content when summarizing its conveyed sentiment. In future work, we additionally aim to investigate the link between sentiment of relevant fragments and sentiment of a text as a whole. Furthermore, we aim to find combinations of sentences constituting a good summary. Last, we plan to implement our findings and to assess different (weighted) combinations of features in order to improve the state-of-the-art in sentiment summarization for conversational text.

Acknowledgments

We would like to thank Teezir (`http://www.teezir.com`) for their technical support, fruitful discussions, and for supplying us with valuable data. The authors of this paper are partially supported by the Dutch national program COMMIT.

References

1. Bal, D., Bal, M., van Bunningen, A., Hogenboom, A., Hogenboom, F., Frasincar, F.: Sentiment Analysis with a Multilingual Pipeline. In: 12th International Conference on Web Information System Engineering (WISE 2011). Lecture Notes in Computer Science, vol. 6997, pp. 129–142. Springer (2011)
2. Beineke, P., Hastie, T., Manning, C., Vaithyanathan, S.: Exploring Sentiment Summarization. In: AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications. pp. 12–15. AAAI Press (2004)
3. Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G., Reynar, J.: Building a Sentiment Summarizer for Local Service Reviews. In: WWW 2008 Workshop on NLP Challenges in the Information Explosion Era (NLPIX2008) (2008), available online, `http://www.cl.cs.titech.ac.jp/~fujii/NLPIX2008/paper3.pdf`
4. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The WEKA Data Mining Software: An Update. SIGKDD Explorations 11(1), 10–18 (2009)
5. Hall, M., Smith, L.: Practical Feature Subset Selection for Machine Learning. In: 21st Australasian Computer Science Conference (ACSC 1998). pp. 181–191. Springer (1998)
6. Lerman, K., Blair-Goldensohn, S., McDonald, R.: Sentiment Summarization: Evaluating and Learning User Preferences. In: 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009). pp. 514–522. Association for Computational Linguistics (2009)
7. Madden, S.: From Databases to Big Data. IEEE Internet Computing 16(3), 4–6 (2012)
8. Mitchell, T.: Machine Learning. McGraw-Hill Series in Computer Science, McGraw-Hill (1997)
9. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval 2(1), 1–135 (2008)
10. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment Classification using Machine Learning Techniques. In: Empirical Methods in Natural Language Processing (EMNLP 2002). pp. 79–86. Association for Computational Linguistics (2002)
11. Pingdom: Internet 2011 in Numbers (2012), available online, `http://royal.pingdom.com/2012/01/17/internet-2011-in-numbers/`
12. Popescu, A., Etzioni, O.: Extracting Product Features and Opinions from Reviews. In: Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT 2005). pp. 339–346. Association for Computational Linguistics (2005)
13. Titov, I., McDonald, R.: A Joint Model of Text and Aspect Ratings for Sentiment Summarization pp. 308–316 (2008)
14. Zhu, J., Zhu, M., Wang, H., Tsou, B.: Aspect-Based Sentence Segmentation for Sentiment Summarization. In: 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion (TSA 2009). pp. 65–72. Association for Computing Machinery (2009)