ESUKA – JEFUL 2025, 16–1: 123–143

TOWARDS AN ESTONIAN DATASET ON DOCUMENT-LEVEL SUBJECTIVITY

Karl Gustav Gailit, Kadri Muischnek, Kairit Sirts
University of Tartu, EE
karl.gustav.gailit@ut.ee, kadri.muischnek@ut.ee, kairit.sirts@ut.ee

Abstract. This article discusses a preparatory step towards developing an Estonian dataset on subjectivity, providing a brief overview of past analyses of subjectivity and the theoretical basis for creating the dataset. Subjectivity has been explored within many fields of linguistics, including pragmatics and formal semantics, as well as natural language processing where computational methods are used to create models for detecting subjectivity, often for further processing. However, many of these models could be improved, and for some it is questionable whether they classify subjectivity or something else, such as text genre. These issues are caused by the datasets these models are trained on, from the text collection method to the unnuanced labels of "objective" and "subjective". To solve this issue, we propose a dataset of documents from various registers with annotations for subjectivity with a scalar value, where zero represents a fully objective document and one a subjective document.

Keywords: subjectivity analysis, subjectivity, datasets, corpora, Estonian, NLP, annotation

DOI: https://doi.org/10.12697/jeful.2025.16.1.05

1. Introduction

Subjectivity is a key aspect of everyday language use since it is important to be able to differentiate between fact and opinion. Because of this, subjectivity is a topic that many fields have touched upon, both in a more epistemological way, such as a general explanation of what subjectivity is and what its characteristics are, and more practically, developing tools and methods for analyzing how it is expressed in texts.

Many aspects of subjectivity have been discussed within these analyses, from extracting subjective parts from texts to examine whether opinions within them are positive or negative to analyzing the functions of lexical features that express subjectivity. Often, these approaches

have differing views on what subjectivity is – some focus in-depth on specific aspects of opinions, such as frequently expressed positive and negative sentiments, while others consider everything that subjectivity and opinion can be.

However, all interpretations and definitions of subjectivity can be simplified into the following sentence: "Subjectivity is the opinion of the author." There are countless opinions, including views and judgments, feelings and emotions, and beliefs. Opinions are contrasted by facts, or objectivity. In general, objectivity can be summarized as "objectivity references reality, not the mind of the author." For instance, saying "it is raining" is objective, but "the weather is terrible" is subjective. However, it should be mentioned that "reality" can differ from author to author: aspects such as religion and societal norms may influence what one considers objective or subjective.

While "subjective" and "objective" are opposites, not all sentences fall into those two categories. Many sentences cannot be classified as either objective or subjective, such as questions and orders. Because of this issue, some discussions involving subjectivity classify sentences into subjective and non-subjective categories, where non-subjectivity involves not only objectivity but also everything outside of the objective-subjective binary. However, this distinction is less necessary for the document level, or analyzing a text in full. A document usually contains multiple sentences, and it is typical for most of them to express some form of subjectivity or objectivity. This, in turn, means that it is likely that the document as a whole contains some form of subjectivity or objectivity, which makes distinguishing additional categories less of a necessity as it is in sentence-level analysis. As this article focuses on subjectivity on the document level, we will use subjectivity and objectivity as the two categories.

In this article, we give an overview of how subjectivity has been analyzed within both linguistics and natural language processing (NLP), focusing on the latter. We discuss how these analyses have been used when creating subjectivity datasets and the problems these datasets contain. We propose a dataset for Estonian, which could remedy these issues by handling subjectivity not as a binary label of objective or subjective but rather as a scale. This dataset aims to provide training material for a model to automatically assign numeric subjectivity values to documents. This value provides insight for further analysis of any

given document, for instance, to determine whether it contains subjective information for opinion mining or to use the value as a feature for detecting whether a journalistic article should be classified as an opinion piece or a news story. This article discusses the preparatory step towards this dataset and the accompanying model.

2. Previous Work

Subjectivity is a topic that can be discussed as an aspect of many fields and areas within linguistics, from more theoretical or philosophical analyses including pragmatics and formal semantics to NLP or analysis of language using computational methods. In this section, we briefly overview how subjectivity has been discussed in various linguistic fields and provide examples of how it has been conceptualized and utilized.

2.1. Subjectivity In Linguistics

As subjectivity falls under the umbrella of non-reality, it is a topic often discussed alongside the realm of modality (Narrog 2012). Modality is an expansive field that primarily examines how language is used to express aspects of truth and reality. It involves topics such as how certain the speaker is about an event happening and how they indicate the basis of their knowledge. However, while subjectivity and modality are often discussed together, it is typically emphasized that subjectivity is an aspect of pragmatics, not modality. For instance, Heiko Narrog (2016) uses modality to provide examples of approaching subjectivity and explicitly states that modality is a grammatical category and subjectivity is considered a part of pragmatics; thereby, the two are independent. Due to this independence, modality can be used to express both subjectivity and objectivity, in contrast to earlier views that consider modality, as the speaker's stance, to be inherently or predominantly subjective.

One way modality can be used to express subjectivity is via epistemic modals. By using modal verbs, the sentence is subjectified, as the speaker brings themselves to the foreground by stating the certainty of their statement, representing the subject's consciousness. The subject

does not need to be the speaker – it can also be someone speaking within the text itself, such as characters in narrative stories. However, the usage of perspective representation expresses non-factuality, with subjectification being relevant due to the subject, the person whose perspective the sentence is uttered, being explicitly in the foreground. (Sanders 1999)

This manner of discussing subjectivity is characteristic to its use in pragmatics, a field of linguistics that examines how context influences meaning during communication, especially within spoken language and conversation. Among other topics, one topic of interest within pragmatics is how one sentence can have multiple meanings and how the context determines which meaning should be interpreted, from the literal meaning of the sentence to how it relates to the previous conversation or the physical world. One of these aspects is subjectivity since it is crucial to know whether the speaker relates a fact or an opinion.

An example of subjectivity being handled within pragmatics is the Estonian project S/IS: "Pragmatics overwrites grammar: subjectivity and intersubjectivity in different registers and genres of Estonian." Subjectivity here is defined as the attitude toward the content that is spoken or written, while intersubjectivity is defined as the attitude toward the conversation partner (Hennoste, Klumpp & Metslang 2022). However, it should be noted that intersubjectivity can also be described as the shared subjective opinion or information of a group as well as the setting itself for the speech situation (Narrog 2016).

Subjectivity is discussed within the project primarily by analyzing various markers and why these specific markers were chosen over others. For instance, an analysis by Helle Metslang (2023) gives a thorough overview of 3 Estonian particles: *ehk*, *äkki*, and *järsku*, where all three of these are translatable into English as 'maybe'. However, these are all polysemic words and have their own nuances in meaning and usage. Corpus-based analysis showed that *äkki* and *järsku* are mainly used intersubjectively, especially to hedge questions and directions within conversations. Meanwhile, *ehk* is primarily used subjectively to express that the sentence contains opinions.

However, as mentioned above, subjectivity in pragmatics is a broad topic. When describing the types of functions of subjectivity, Hendrik De Smet & Jean-Christophe Verstraete (2006) highlight pragmatic subjectivity, where the speaker's perspective is the basis for the selection of

expression used depending on how they interpret reality. For instance, describing a person as *honest* or *short* expresses how the speaker interprets honesty and height. They contrast this type of subjectivity with two types of semantic subjectivity, where subjectivity is inherent to the meanings of expressions. The first is ideational subjectivity, which handles the extralinguistic and extradiscursive world or the speaker's subjective attitude about the situation. It is expressed by adjectives such as *pleasant* and *horrifying*, which show the speaker's positive or negative stance, unlike pragmatic subjectivity. The second is interpersonal subjectivity, or how the speaker positions themselves in relation to their interaction with their conversational partner.

While subjectivity is a popular topic in semantics, one branch that discusses it thoroughly is formal semantics. As a field of study, formal semantics handles language as a set of mathematical rules and equations. It analyzes how words can be composed into sentences and how this composition influences the meanings of the sentences. One aspect of formal semantics that originates from mathematics is the truth condition of the sentence, where whether the sentence is true or false depends on both the meanings of the individual words within the sentence and how the sentence relates to reality.

A rule here is that a sentence can either be true or false and that if a true sentence is negated, it becomes false, and vice versa. For instance, negating the true sentence "The sky is blue" results in the false sentence "The sky is not blue." However, this rule does not apply in all situations, including subjective sentences, where it results in faultless disagreement. This means that a sentence and its negation are simultaneously true. Compare these two pairs of sentences:

– "This Picasso painting is beautiful."
  "No, it's not beautiful."
– "This chair is four-legged."
  "No, it's not four-legged."

The sentences in the first pair can both be simultaneously true, whereas the sentences in the second pair contradict each other, meaning one of these sentences must be true and the other false. In both examples, the adjectives create the focus of subjectivity analysis, where *beautiful* is categorized as subjective and *four-legged* as objective.

However, there exists a third category for adjectives, a mixed category, where, depending on the context, an adjective can be either objective or subjective (Solt 2018). One such adjective is *high*. For instance, the negation of the sentence "Suur Munamägi is a high peak" has a context-dependent truth value: while Suur Munamägi is Estonia's highest peak, it has a height of only 318 meters and, in comparison to the highest peaks of other countries, it can be considered to be a low peak. Even color terms, which are generally considered objective adjectives, can cause disagreements based on subjective opinions, such as whether the color of oxidized copper or the sea is blue or green.

Additionally, the analysis of subjectivity is not limited to adjectives. Fleisher (2013) describes the existence of two different types of subjectivity to explain why the vague predicate *find* cannot be used in the same contexts as taste predicates, such as *consider*. The types of subjectivity described are "mapping subjectivity" and "standards subjectivity". Mapping subjectivity occurs when different speakers can map a specific word or phrase to different values on a scale, for instance, how beautiful a painting is. However, adjectives like *high* map to the height of the described object, thus lacking mapping subjectivity. Standards subjectivity, however, looks at the value and decides whether it is up to the speaker's subjective standards or not, such as the cutoff value for beauty for something to be considered beautiful and the cutoff value for height for something to be considered high. While taste predicates require the existence of either standards or mapping subjectivity, the vague predicate *find* is only sensitive to the existence of mapping subjectivity and thus can only be used with it.

2.2. Subjectivity in Natural Language Processing

Discussions on what subjectivity is and what means are used to present it are also a matter of discussion within the realm of natural language processing (NLP), where the task of subjectivity analysis exists. The goal of this task is to use computational methods to analyze text to gain a better understanding of subjectivity as a linguistic phenomenon, as well as facilitate further subjectivity analysis methods for both linguistic knowledge as well as the practical applications subjectivity analysis offers.

While NLP typically focuses on the practical applications of subjectivity analysis, there have also been some more theoretical analyses within this field. After all, to accurately analyze and predict subjectivity, one has to first understand what subjectivity is and how it is expressed. Knowledge of subjectivity is important to create the models themselves, to better understand their predictions, and to see what aspects of subjectivity cause problems for the model. Wei Chen (2008) points out six aspects, or dimensions, of subjectivity that one must consider when discussing the topic. These six aspects are 1) non-objectivity or opinions, 2) uncertainty, 3) vagueness, referring to concepts without exact definitions including gradable adjectives like *high*, 4) non-objective measurability, where adjectives such as *beautiful* are not possible to be measured without bias, 5) imprecision, or the agreement that numerical values do not need to be exact, such as saying someone is "two meters tall" despite their height varying slightly from exactly 2000 millimeter, 6) and ambiguity, where a statement can have multiple meanings and explanations.

The more theoretical analysis of subjectivity has provided a general methodology and the necessary tools for the task of subjectivity analysis. As time has passed, the tools and methods have been developed further to utilize newer technological advances, the most recent example being the use of large language models. The theoretical aspect, however, has mostly remained as it was. At best, it is used as a reference when the articles describing these more practical developments need to provide a brief overview of what subjectivity is as a concept. In the current day, subjectivity analysis is typically part of a more extensive operation where a simplified approach is used, with texts, either documents or sentences, being labeled as 'objective' or 'subjective'. Additional labels may also be used to denote sentences that do not fit under those two labels, such as questions, or to mark the sentences that the annotators were not able to label. These labels cannot provide more in-depth analysis, unlike the linguistic analyses discussed in section 2.1. However, this simplified approach is standard in NLP since not only is it less complex and thus faster to annotate, but it also provides data suitable for further computational analysis.

As a sidenote, the terms "text", "document", and "sentence" are often used in NLP. "Text" is the broadest of these, as any instance of written language is considered a text. The more traditional meaning

of a complete written work is under the term "document", including, for instance, full articles, forum threads, recipes, and posts from microblogs. The definition of a "sentence" is the same as in linguistics: a set of words that together express a thought. Both sentences and documents are types of text. It should be noted that NLP often discusses text sourced from the internet, where the scale of a document may be very small. For instance, a single review in an online store is a document, even if said review is made up of a single sentence or word.

While computationally analyzing subjectivity can be viewed as an individual task, it is more often considered a subtask of sentiment analysis and opinion mining. These two tasks analyze opinionated sentences or phrases, which first have to be identified using subjectivity analysis. Due to this, both sentiment analysis and opinion mining fall under the umbrella of subjectivity (Liu 2020), with several studies and datasets on these subtopics being marked to be about subjectivity in general. This approach is harmful, as it causes ambiguity in understanding what task is meant when the keyword "subjectivity" is used. Subjectivity detection as a standalone task involves assigning sentences or, less commonly, larger sections (including whole documents) "objective" and "subjective" labels.

For example, Janyce Wiebe & Ellen Riloff (2005) used rule-based sentence subjectivity classification models to create a dataset containing subjective and objective sentences. They created two models, one to find subjective sentences and one to find objective sentences. These models were constructed manually, not using machine learning, and the rules within these models were constructed based on prior subjectivity-related literature and linguistic knowledge. These rule-based models were tested on sentences from the Multi-Perspective Question Answering Corpus and evaluated against the subjectivity labels within the corpus. Both the subjectivity and objectivity detection models had high precision but low recall, meaning while the models did not assign these labels to sentences that should not have had labels, they also did not label many of the sentences that were supposed to be labeled. This shows that while the rules used within these rule-based models could accurately represent a section of subjectivity, not the entirety of it.

A brief note on the use of the terms *corpus* and *dataset*. A dataset is any collection of data, and as such, its contents and format can vary, while a corpus is a type of dataset that contains linguistic data, a

collection of written texts. The texts within corpora can be full documents or smaller sections, for example sentences. They may contain additional data related to these texts, such as subjectivity labels, or consist exclusively of texts. In NLP, these terms are often interchangeable, as the data is linguistic in nature. When referring to previous works, this article uses the same terms the authors themselves have used, whether it be corpus or dataset.

Another rule-based method of analyzing sentence subjectivity is analyzing the subjectivity of the individual words within the sentence. This is how TextBlob,[1] a popular Python library for subjectivity and sentiment analysis, scores sentence subjectivity. TextBlob has an internal lexicon where every word has a subjectivity value between zero and one, where zero is objective and one is subjective. When calculating the subjectivity of a sentence, only those words that exist in the lexicon are considered. The subjectivity value of the sentence is the mathematical mean of the words within the sentence. This method, however, is very limited by the contents of the lexicon, with a big issue being new or uncommon words that do not exist within the lexicon and thus have no assigned subjectivity value. Therefore, these words do not impact the score of a sentence as one would assume, giving an incorrect score. Additionally, this method is heavily influenced by words with multiple meanings, where a word is objective in one context and subjective in another. To resolve that issue, TextBlob assigns subjectivity scores to the appropriate WordNet synset, a group synonyms or words that share the exact same meaning, rather than the word itself. This allows for differentiation between the meanings of a word, with each meaning having a separate score. In the case where it is unclear which meaning is used for a given word, TextBlob uses the mean score of all of the synonyms.

However, machine learning-based methods are now far more common than rule-based models. For instance, Pavel Přibáň & Josef Steinberger (2022) analyzed subjectivity in Czech sentences sourced from movie reviews and plot summaries. The expectation was that reviews would mostly consist of subjective sentences and summaries mostly of objective sentences. They had two human annotators review these sentences and classify them as objective or subjective. The annotators used additional tags to denote sentences that were either questions,

1 TextBlob homepage: https://textblob.readthedocs.io/en/dev/ (Last visited 17.04.2025)

unanalyzable due to errors in their structure, or which they were unsure about. Based on a total of 11,907 analyzed sentences, reviews comprised 72.6% of subjective sentences and 11% of objective sentences. Meanwhile, summaries comprised 84.2% objective sentences and 3.9% subjective sentences. This human-annotated data was supplemented by 100,000 sentences from movie reviews marked as subjective and 100,000 sentences from movie descriptions marked as objective. These were used to train multiple models; the best model was based on XLM-R-Large and had an accuracy of 93.56%. Using only supplementary data without the human-annotated sentences, the accuracy of this model dropped down to 91.96%.

While less common than sentence-level analysis, subjectivity analysis can also be performed on the document level. However, this analysis is typically domain-specific, with the most used method of labeling marking the label of the document based on its source. For example, in the domain of newspaper articles, opinion pieces would be labeled subjective, while news would be labeled as objective. This method was used to create datasets in the 2009 DEFT shared task, predicting whether a newspaper article is subjective or objective (Toprak & Gurevych 2009).

It should be noted that there are many methods used to train and develop document-level subjectivity analysis models. The most common method is to handle a document just like a sentence in the sentence-level analysis models, with the document being a singular string with a subjectivity value. However, many methods of training NLP models are limited by input size, making it difficult to input a whole document and requiring internal segmentation. Another method to score subjectivity on the document level is scoring smaller sections, typically sentences, and calculating the mean value of these sections to get the score for the entire document.

Subjectivity analysis has additionally been used within other projects where it is important to detect whether texts are opinionated. One such topic is bias detection. According to Marta Recasens, Cristian Danescu-Niculescu-Mizil & Dan Jurafsky (2013), subjectivity is the cause of a type of bias called framing bias, which is expressed by phrases and words that express a specific, subjective point of view, for instance showing distaste towards a type of large house by calling them McMansions. The goal of their trained model is to identify the word

that causes bias in each sentence. While the model did achieve similar results to human annotators, it should be noted that this identification is a difficult task, with the model's accuracy of 34% being close to the human accuracy of 37%.

3. Comparing subjectivity between texts

While the methods used to create the datasets mentioned above lead to usable datasets that fulfill their intended purposes, it is undeniable that the methods used are imperfect and contain many issues, some more critical than others. The most prominent issue is the common usage of source-based subjectivity assessment. While it is true that some text genres and subgenres are inherently subjective or objective, for instance opinion pieces and news articles, it is not the case for all genres. Additionally, by creating a dataset based on sources rather than the content of the texts themselves, the following question arises: does the dataset reflect subjectivity, or is it the genres of the texts that do? Compare the product reviews and product descriptions on online store – reviews should be subjective, and descriptions should be objective. While this is the case with most texts, there are still many exceptions, especially on the sentence level.

Take a sentence that belongs in a hotel review: "The room had a view of the pool." This sentence is objective since it only references the real world and does not contain any of the author's thoughts. However, since it is from a review, a model trained on data where subjective texts are made up of reviews will label this sentence as subjective. This label is caused by the model recognizing that the structure of this sentence is very similar to those typically found within reviews, for instance, the usage of the past indefinite tense. As reviews and descriptions differ in language use, including sentence structure, common words and phrases and even tense, most models will recognize these differences during training and in turn learn to misclassify the sentences such as the one mentioned above as subjective. This issue, where the models learn to classify based on the genre features, is present in most models trained on datasets created using the source-based method, making it a problem with the method itself.

There is also the issue of ambiguity when it comes to understanding where a sentence may be sourced from. A sentence "This is a wonderful blend of well-aged tea and flavorful fruits" is subjective, which is apparent due to the usage of opinions and the author's sensory experiences. However, this does not mean that this sentence was undoubtedly from a review since many product descriptions contain similar sentences to evoke feelings and emotions in potential customers. Take, for instance, an Estonian product description and its translation (1).

(1) Meie teekollektsiooni vaieldamatu pärl!
Fu Lu Gongi nimelisest teevabrikust Feng Shanist, Hiinast.
Meie tee on korjatud 1990. aastal ning on hoiustatud niiskes ja soojas Malaisias, kus seda peeti puuduvate paberümbriste tõttu pikka aega ekslikult kohalikuks teeks, avastades tee tegeliku päritolu alles aastaid hiljem.
Suurepärane toormaterjal iidsetelt metsikult kasvavatelt teepuudelt.
Hoiustamise tingimuste tõttu on tee laagerdunud kiiresti, ent väga kaunilt.
Soovitame juua seda teed meditatiivselt. (Kallas & Koppel 2022, Web 2019 corpus, id:4244262)

(Translation of example 1) The undisputed gem of our tea collection!
From the Fu Lu Gong tea factory in Feng Shan, China.
Our tea was harvested in 1990 and has been stored in humid and warm Malaysia, where it was mistaken for a local tea due to missing paper wrappers for a long time, with its true origins discovered only years later.
Excellent raw material from ancient wild-growing tea trees.
Due to the storage conditions, the tea has matured quickly but very beautifully.
We recommend drinking this tea meditatively.

Example 1a is a product description of a tea imported from China. The language used within this text is very illustrative and descriptive, containing many phrases that are generally considered subjective, such as claiming that the tea has aged "beautifully" and that it is the "undisputed gem" of the store's collection of tea. Additionally, there is a recommendation to drink the tea "meditatively", which is a very subjective suggestion. The text also contains objectivity that is more typical of product descriptions, including a description of where and when this tea was harvested and how it was stored, providing information on what

this product is. However, this text should not be labeled as purely objective, as its genre, product description, suggests.

Both product reviews and descriptions contain a large number of phrases and words unique to them, which machine learning models will detect and use for learning. However, many of these domain-specific words and phrases are not usable for subjectivity analysis. For instance, product reviews often discuss how a product was delivered, e.g. product arriving broken, but product descriptions do not discuss this topic. Therefore, the sentence "The product arrived broken", belonging most likely to a product review, would be classified as subjective despite it being objective since it reflects reality. Therefore, the models trained on source-based data learn to classify the sources they were trained on rather than label subjectivity and objectivity, which results in a model that ultimately classifies the texts into the two source genres rather than objective or subjective.

Another common issue with classifying document subjectivity is the variability of subjectivity within the sentence level. It is rare for a document to only contain objective or subjective sentences: opinion pieces typically contain facts to support their opinions, and news often have interview sections where someone expresses their opinion on a certain topic, especially within the realm of politics. While this variability was noted within the analysis of subjective and objective texts by Přibáň & Steinberger (2022), it was not the first time this was noted. Janyce Wiebe, Theresa Wilson & Matthew Bell (2001) had noted before that in journalism, opinion pieces contained 70% subjective sentences and 30% objective sentences, and the ratio was smaller for non-opinion pieces, with 56% of sentences being objective and 44% subjective.

This evokes a question: what is it that makes a document subjective? Is it simply a matter of quantity, and the subjectivity of the entire document depends on the fact whether there are more subjective or objective sentences within it? This might be the case for most documents, but there will always be exceptions. For instance, an opinion piece on nuclear energy might be structured in a way where the author's opinion is supported by multiple facts and scientific research, yet the document itself is still an opinion piece. While there might be two or more facts for every opinion in that text, it is still clear that the document itself is subjective since the goal of this text is to convey an opinion.

Therefore, is the intention of the author the sole arbiter of the document's subjectivity? Again, this is the case for most documents. However, there will always be cases where the intention of the author is not clear to the reader – for example, is a document an opinion piece with too many facts, or is it a news article that contains too many opinions? In these cases, the author's intention may not align with the intention the reader understood from the document itself, meaning that an opinion piece can be understood to be objective, or vice versa, a news article can become subjective. While this sort of interpreting is something people typically do well, machines will struggle to understand the intention behind a document. Additionally, some documents are not written to inform the reader of facts or to spread the author's opinion, for instance, fiction and poetry are mainly written to evoke emotions within the reader, which is a different goal from the author expressing their own feelings regarding a topic. Instructions, such as recipes or manuals, aim to instruct the reader how to perform a specific action, not provide them with objective facts.

It is likely that document subjectivity is not caused by the subjectivity of the sentences themselves or the intention with which the author wrote them, but rather, they are simply aspects of what makes a document subjective or objective. And while both of these aspects are important for figuring out the subjectivity of a document, reading and understanding a document is a complicated process, and there may be additional factors that make a document read as subjective or objective.

4. Suggesting the scalar subjectivity dataset

It is possible that a dataset which labels the data within it as either subjective or objective does not contain suitably nuanced information for training accurate subjectivity analysis models. One option to solve this problem is to expand this binary label into a scale of numeric values. While handling subjectivity and objectivity as a numeric scale is not a typical approach, there has been previous research where the strength of subjective elements has been annotated. When having annotators identify subjective elements in a sentence, Wiebe (2000) also had each of the elements scored on how strongly subjective they were. While the scale only spanned from one to three, with three being the strongest, this still

serves as a basis for subjectivity as a numeric value. Louis Escouflaire, Antonine Descampe & Cédrick Fairon (2024), when having annotators assess the subjectivity of excerpts from Belgian journalistic texts, used a Likert scale with five points, one being a completely objective text and five an entirely subjective text. Since the excerpts were annotated by multiple annotators, the subjectivity score of an excerpt is the mean of the annotations, therefore using the full numeric scale between one and five.

However, with the annotators only having a limited number of options, many nuances could go missing. Expanding these small ranges into a full numeric scale would provide more detailed and accurate information. Therefore, we propose that the next step is to make a dataset of documents that have been annotated with their subjectivity values as scores between zero and one. In it, zero represents complete objectivity, one represents complete subjectivity, and values between zero and one represent more nuanced documents that contain aspects of both subjectivity and objectivity. These values are assigned by human annotators using their innate ability to comprehend the documents and their meanings, including their subjectivity, from the intention of the document to how it conveys the opinions and facts within it. Additionally, after scoring a document, the annotators will not be asked to explain their reasoning. Since the goal is to annotate intuitively, asking for reasoning will prime annotators to look for said reasoning, leading them to focus on aspects of subjectivity already known to them, such as subjective adjectives and epistemic verbs.

It is important to note that these annotations will be inherently subjective since every annotator can score a single document differently. To mitigate this issue, every document should be scored by multiple annotators, and the assigned subjectivity value would be the average of their scores, similar to Escouflaire, Descampe & Fairon (2024). This will help prevent bias annotators as individuals have, and the dataset will reflect a more general opinion. Additionally, data from the individual annotators can be used to gain further information, such as calculating inter-annotator agreement, making it possible to analyze which documents get more varied subjectivity scores and thus are more challenging to annotate.

While it is possible to mitigate the issue of subjective and varied scores by giving the annotators detailed instructions and rules on how

to score the documents, this additional guidance also removes the aspect of intuitive subjectivity assessment. Training the annotators makes for a cleaner and more stable dataset for machine learning, but it also means that this dataset has a prescribed meaning of subjectivity based on previous research. Thus, it cannot be used to find and analyze additional potential aspects humans use to detect and recognize subjectivity and objectivity. Additionally, including intuitive subjectivity assessment allows for analysis regarding other aspects of text, such as how subjectivity relates to the writing style, genre, or topic.

It should be noted that the scalar values within this dataset can be easily translated into binary labels, thus making it usable for applications requiring only objective and subjective labels. Values below an assigned threshold can be assigned the objective label, while values above the threshold can be assigned as subjective. These thresholds can be assigned by the user, allowing for versatility on what they consider as subjective and objective, thus allowing them not to keep documents that do not suit their needs.

Additionally, a similar method could be used to group the documents within the dataset by their subjectivity scores, similar to Escouflaire, Descampe & Fairon (2024) and Wiebe (2000). While less informative than using the scores themselves, this method would still provide a way to compare different intensities of subjectivity. It would mitigate the issue of potential ambiguity between similar scores, where it could be unclear how small differences in score influence how strongly subjective the document can be read. However, this issue does not influence how this dataset could be used to train models due to the additional level of detail provided by the numeric annotations.

To further explain how handling subjectivity as a scale could be useful, we provide a couple of examples (2, 3).

(2) Päris kindlasti pole ma kuulnud gruuvivamat ja paremini produtseeritud Eesti rockiplaati kui seda on Revolver II. Igas mõttes viimase peal. Toimib igasuguse tilulilu ja pseudospetsialistide haibita. Korralik, munadega kitarrimäng ja õige hoiak. Rock on vibratsioonivärk, seda teeselda ei saa... ükskõik kui ilus poiss sa oled. Jaanus Nõgisto

Oskus segada vanu klassikuid ja uusi suundi, kohalikke ja välismaised, lisadest törtsu põhjamaist viha ja ärritatust maailma vastu, mis tuleb vaid läbi tule ja vee käinud rahvuslusest. Ühesõnaga on poisid suutnud ühendada terve rockižanri parimad palad, suutes jätta veel ruumi

kodumaiselt tuttavale soojale sarkasmile, mis kõlab nii lüürikas kui ka muusikas. Joonas Tepp, KASSETT. EE (Kallas & Koppel 2022, Web 2017 corpus, id:1674701)

(Translation of example 2) I've definitely never heard a groovier and better-produced Estonian rock record than Revolver II. It's the best in every way. Works without any frills and pseudo-specialist hype. Good, ballsy guitar playing and the right attitude. Rock is a vibe thing, you can't fake it… no matter how pretty a boy you are. Jaanus Nõgisto

The skill to mix old classics and new trends, local and foreign, adding a bit of northern anger and irritation towards the world, which only comes from nationalism that has been through fire and water. All in all, the boys have managed to combine the best songs of the whole rock genre, while being able to leave room for warm sarcasm familiar from home, that can be heard in both the lyrics and the music. Joonas Tepp, KASSETT. EE

(3) „450–500 kr eest soovitaksin osta pigem mingid teised klapid kui need sony-d (Sennheiser). Olen üsna pettunud heli kvaliteedis. Ülemised sagedused domineerivad ning alumine bassi ots kohati nagu täitsa puudub, eriti alla 50Hz. Juhe on oma 3 meetrit pikk ja tüütu. Väliselt jätavad natuke odava mulje, enamuse klappidest on pastik [*sic*!]. Peas on muidu mugavad ning kerged. Sobivad vanaemale raadio kuulamiseks, kuid korralikku heli hindavale inimesele jääb neist ilmselgelt väheks." (Kallas & Koppel 2022, Web 2019 corpus, id:3756491)

(Translation of example 3) "For 450–500 kr, I would recommend buying some other headphones rather than these Sony ones (Sennheiser). I am quite disappointed with the sound quality. The upper frequencies dominate, and the lower end of the bass is sometimes completely absent, especially below 50Hz. The cord is 3 meters long and annoying. Appearance-wise, they give a bit of a cheap impression, most of the headphones are plastic. They are otherwise comfortable and light when worn. Suitable for a grandmother to listen to the radio, but for a person who appreciates good sound, they are obviously not enough."

Both documents originate from reviews and can generally be considered subjective. However, they differ from each other in most aspects. The document in example 2a consists of two reviews of an Estonian rock album. The reviewers sing its praises and talk about the

album's "vibe" and how skillful the band is. Clearly, all the sentences in this document are the reviewers' opinions, and not a single sentence can be considered traditionally objective. The document in example 3a is a review of a pair of headphones. It is a negative review with many opinions, calling the long cord "annoying", the appearance "cheap", and the sound quality "disappointing". However, this review provides reasoning on why the author of the review thinks this way, something the reviewers in document 2a did not do. The cheap appearance is backed by the headphones being made of plastic, and the sound quality concerns are expressed by analyzing how these headphones emit sound frequencies.

Comparing both examples, it is clear that while both of these documents are subjective, their subjectivity can be contrasted. Due to the lack of explanations provided within document 2a, it reads as noticeably more subjective than document 3a. Within the proposed dataset, the human annotators would likely annotate document 2a very highly, for instance the maximum value of 1, while document 3a will have a slightly lower score, such as 0.85; however, the difference between the subjectivity of these two documents is clear. Additionally, there might be other factors than the proportion of justification that influences how people interpret subjectivity. By providing a dataset that contains documents with comparable, numeric scores of subjectivity, it would be possible to conduct a more fine-grained linguistic analysis. For instance, this dataset could be used in the field of pragmatics to analyze which phrases and other linguistic features correlate with higher and lower subjectivity scores, which, in turn, would allow for a better understanding of subjectivity itself.

5. Conclusion

Subjectivity is a complicated topic, not only for machines but also for humans. Within linguistics, it has been analyzed how people express opinions and facts and what specific methods are used to differentiate those two. Subjectivity is a topic of research in multiple fields of linguistics, including pragmatics, modality, and formal semantics, but it is also important to natural language processing (NLP). Knowledge, whether something is objective or subjective, is essential for text

analysis, especially for the task of sentiment analysis – to know if a sentence contains positive or negative opinions, it is important to know whether the sentence contains any opinions in the first place.

While many NLP models and studies on predicting subjectivity exist, advancements can still be made in this area of research. Rule-based models, while based on linguistic theory, tend to lack nuance since they tend to avoid tagging more ambiguous instances as neither objective nor subjective. However, models based on machine learning are limited by the datasets on which they have been trained. Many datasets tend to be domain-specific, containing data only from certain types of texts, such as opinion pieces for subjective texts and news for objective texts. However, even cross-domain datasets are limited by their lack of nuance within the binary labels of "objective" and "subjective". While *opinion* and *fact* are separate concepts, there does exist ambiguity between these two opposite ends, which becomes more prominent the larger the analyzed piece of text is, from phrases and sentences to whole documents. Because of this ambiguity, many texts can be compared to one another, with one text being more subjective or objective than the other.

To mitigate this issue, we proposed a subjectivity dataset consisting of documents labeled with a numeric value between zero and one, where zero is a completely objective document, one is a completely subjective document, and numbers between those two express a more nuanced level of subjectivity between the two extremes. Since each document within the dataset is evaluated by multiple annotators, it will be possible to use the human ability to understand and interpret subjectivity while reducing the issue of annotator bias. The dataset will, therefore, contain more nuanced data on subjectivity, which, in turn, supports the development of more accurate and realistic models for analyzing subjectivity.

References

Chen, Wei. 2008. Dimensions of subjectivity in natural language. *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies Short Papers*. Columbus, Ohio: Association for Computational Linguistics. https://doi.org/10.3115/1557690.1557695.

Escouflaire, Louis, Antonin Descampe & Cédrick Fairon. 2024. Unveiling Subjectivity in Press Discourse: A Statistical and Qualitative Study of Manually Annotated Articles. *Discours* 34. https://doi.org/10.4000/12jg5.

Fleisher, Nicholas. 2013. The dynamics of subjectivity. *Semantics and Linguistic Theory* 23. 276–294. https://doi.org/10.3765/salt.v23i0.2679.

Hennoste, Tiit, Gerson Klumpp & Helle Metslang. 2022. Diskursusemarkerid ja pragmaatika. Sissejuhatuseks. *Keel ja Kirjandus* 65(1–2). 3–18. https://doi.org/10.54013/kk770a1.

Kallas, Jelena & Kristina Koppel. 2022. *Estonian National Corpus 2021*. Center of Estonian Language Resources. https://doi.org/10.15155/3-00-0000-0000-0000-08D17L.

Liu, Bing. 2020. *Sentiment analysis: mining opinions, sentiments, and emotions. Studies in Natural Language Processing* [2nd edition]. Cambridge, New York: Cambridge University Press. https://doi.org/10.1017/9781108639286.

Metslang, Helle. 2023. Kolme (inter)subjektiivsusmarkeri lugu: *ehk*, *äkki*, *järsku*. *Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics* 14(2). 271–296. https://doi.org/10.12697/jeful.2023.14.2.11.

Narrog, Heiko. 2012. *Modality, Subjectivity, and Semantic Change: A Cross-Linguistic Perspective*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199694372.001.0001.

Narrog, Heiko. 2016. Three types of subjectivity, three types of intersubjectivity, their dynamicization and a synthesis. In Daniel Olmen, Hubert Cuyckens & Lobke Ghesquière (eds.), *Aspects of Grammaticalization*, 19–46. De Gruyter. https://doi.org/10.1515/9783110492347-002.

Přibáň, Pavel & Josef Steinberger. 2022. Czech Dataset for Cross-lingual Subjectivity Classification. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi et al. (eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France*. 1381–1391. European Language Resources Association.

Recasens, Marta, Cristian Danescu-Niculescu-Mizil & Dan Jurafsky. 2013. Linguistic Models for Analyzing and Detecting Biased Language. In Hinrich Schuetze, Pascale Fung & Massimo Poesio (eds.), *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Sofia, Bulgaria,* 1650–1659. Association for Computational Linguistics.

Sanders, José. 1999. Degrees of subjectivity in epistemic modals and perspective representation. In Leon De Stadler & Christoph Eyrich (eds.), *Issues in Cognitive Linguistics*. Berlin, New York: DE GRUYTER MOUTON. https://doi.org/10.1515/9783110811933.471.

Smet, Hendrik De & Jean-Christophe Verstraete. 2006. Coming to terms with subjectivity. *Cognitive Linguistics* 17(3). 365–392. https://doi.org/10.1515/COG.2006.011.

Solt, Stephanie. 2018. Multidimensionality, Subjectivity and Scales: Experimental Evidence. In Elena Castroviejo, Louise McNally & Galit Weidman Sassoon (eds.), *The Semantics of Gradability, Vagueness, and Scale Structure. Language Cognition, and Mind 4*, 59–91. Springer International Publishing. https://doi.org/10.1007/978-3-319-77791-7_3.

Toprak, Cigdem & Iryna Gurevych. 2009. Document level subjectivity classification experiments in deft'09 challenge. *Actes du cinquième DÉfi Fouille de Textes*, 22.06. 91–99.

Wiebe, Janyce M. 2000. Learning Subjective Adjectives from Corpora. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Wiebe, Janyce & Ellen Riloff. 2005. Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. In Alexander Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing. Lecture notes in Computer Science 3406*, 486–497. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-30586-6_53.

Wiebe, Janyce, Theresa Wilson & Matthew Bell. 2001. Identifying Collocations for Recognizing Opinions. *Proceedings of ACL/EACL 2001 Workshop on Collocation, Toulouse, France*.

Kokkuvõte. Karl Gustav Gailit, Kadri Muischnek, Kairit Sirts: Eestikeelse terviktekstide subjektiivsuse andmestiku suunas. Artikkel selgitab eestikeelse subjektiivsuse andmestiku loomise ettevalmistavat sammu, andes seejuures ülevaate varasematest subjektiivsuse käsitlustest ning teoreetilise aluse andmestiku koostamiseks. Subjektiivsust on käsitletud mitmes lingvistika harus, sh pragmaatikas ja formaalses semantikas, aga ka loomuliku keele töötluses, kus kasutatakse arvutuslikke meetodeid, et luua mudeleid subjektiivsuse tuvastamiseks, mille eesmärk on sageli andmestiku edasi töötlemine. Mitut sellist mudelit on võimalik edasi arendada, mitme puhul tekib aga küsimus, kas need klassifitseerivad subjektiivsust või midagi muud, näiteks žanrit. Probleemid on tingitud andmestikest, mille peal on mudelid treenitud, kuidas tekste on kogutud ning sellest, et sildid „objektiivne" ja „subjektiivne" on jäigad. Nimetatud probleemide lahendamiseks pakume välja andmestiku, mis sisaldab tekste mitmest registrist ning mis on märgendatud arvuliste subjektiivsuse hinnangutega, kus null tähistab objektiivset teksti ning üks subjektiivset teksti.

Märksõnad: subjektiivsuse analüüs, subjektiivsus, andmestikud, korpused, eesti keel, NLP, märgendamine