Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2019”

Moscow, May 29—June 1, 2019

AN ANATOMY OF A LIE: DISCOURSE PATTERNS IN ULTIMATE DECEPTION DATASET

Pisarevskaya D. (dinabpr@gmail.com)

Institute for Systems Analysis FRC CSC RAS, Moscow, Russia

Galitsky B. (boris.galitsky@oracle.com)

Higher School of Economics, Moscow, Russia and Oracle Corp, Redwood Shores CA, USA

We propose a hypothesis that a deception in text should be visible from its discourse structure. The problem of deception detection is then formulated as classification of a discourse tree of this text, according to the Rhetorical Structure Theory. This discourse tree (DT) is extended by the speech acts expressions attached as the labels for the edges. We employ what we call an ultimate deception dataset: a set of customer complaints for English, that includes descriptions of problems customers experienced with certain businesses. It contains about 2,400 complaints about banks and provides clear ground truth, based on available factual knowledge in the financial domain. The complaints are written by non-professional writers. We conduct experiments to explore correlation between implicit cues of the rhetorical structure of texts and how truthful/deceptive are these texts. The results show that a deception in text can be detected reliably enough to assure industrial applications. Automated detection of text with misrepresentations such as fake reviews is an important task for online reputation management.

Keywords: Customer Complaints, Rhetorical Structure Theory, Discourse analysis, Deception detection, Fake Reviews

1. Introduction

It has been discovered that a lot of forms of human intellectual and communication activity are associated with certain discourse structures. Rhetorical Structure Theory (RST) [1] is a good means to express correlation between such form of activity and its representation in how associated thoughts are organized in text. Rhetorical Structure Theory presents a hierarchical, connected structure of a text as a discourse tree, with rhetorical relations between its parts. The smallest text spans are called elementary discourse units (EDUs). In communicative discourse trees (CDTs), the labels for communicative actions (CAs) (VerbNet expressions for verbs) are added to the discourse tree edges to show which speech acts are attached to which rhetorical relations; this structure helps to understand argumentation [2].

Logical Argumentation needs a certain combination of rhetorical relations of *Elaboration, Contrast, Cause* and *Attribution* [3]. Persuasiveness relies on certain structures linking *Elaboration, Attribution* and *Condition* [4]. Explanation needs to rely on certain chains of *Elaboration* relations plus *Explanation* and *Cause*. A rhetorical agreement between a question and an answer is based on specific mappings between the rhetorical relations of *Contrast, Cause, Attribution* and *Condition* between the former and the latter [5]. Discourse trees turned out to be helpful to form a dialogue and to build dialogue from text, in order to better understand the structure of texts.

In this paper, we study rhetorical structure correlated with certain forms of verbal activity, namely we focus on deception in texts of various genres such as news articles, customer reviews and customer complaints. We intend to discover the distinct features of discourse trees associated with deception. Some of such features can be observed as a result of manual analysis, but most of such features are concealed and need to be tackled by a data-driven approach, so we adjust our customer complaints dataset tagged to detect improper argumentation patterns and invalid claims to serve as a training/test dataset for detection of deceptions.

Research on automated deception detection in written texts is focused on classifying if a narrative is truthful or deceptive. Even if an exhaustive factual information / ontology for a domain is available, it is still hard to perform fact-checking in texts since substantially deep text understanding is necessary and text representation via a logic form is required. It is much more difficult to assess truthfulness when such ontology does not exist, as even manual deception detection, in order to collect datasets for machine learning, as a biased and subjective task. The main difficulty is to detect deception where factual knowledge is not available to a degree sufficient to computationally establish the truth. This situation is typical in the real world, from intuitive choice of product based on reviews to judges' verdicts. It is impossible to establish the truth based on known facts, so decisions are based on implicit cues such as the way people explain what they have done and provide arguments for why they have done so.

While detecting misrepresentation in writing, it is important to differentiate between different categories of writers. Professional writers are frequently good at misrepresenting, and they do not include cues for what might be a lie. Conversely, a content written by non-professional writers is often authentic in how it indicates the thought patterns of the writer where the traces of a lie and hints for how it is motivated can be found.

That's why we analyze how misrepresentation occurs in both professional writing and user generated content (and provide examples of different genres: customer complaints and news stories). Due to this reason, we also provide the ground truth dataset that contains texts written by non-professional writers (bank customers). We also evaluate our classifier, trained on the new dataset, in the domain of business correspondence of non-professional writers such as Enron dataset.

We focus on deception in reviews of products and services as a special case. Automated detection of fake reviews is important for online reputation management tasks. Since fake reviews dataset is available, this is a good domain to evaluate our general domain-independent deception detection algorithm. Fake reviews are deception, but they are artificial since their purpose is not to do a misrepresentation to achieve an agent goal. Usually, this goal is associated with a desired action of another agent who is the addressee of the text that includes this misrepresentation (that is a main scenario of why people lie in the real world). Instead, in the domain of reviews, its subgenre—fake complaints—are written on demand to manipulate public opinion, that is not an usual purpose of misrepresentation in interaction between people expressed in text. They are written with a definite objective, in order to get a better service after the complaint. Therefore, we believe that customer complaints could be the most adequate data source to explore the linguistic correlates of deception and train a classifier.

In customer complaints, complainants frequently write that they have been provided a misrepresentation by a customer support personnel. At the same time, it might be possible that the complaints are in turn lying about what was said to them by their opponents. It is hard to determine, who is lying: customer support or the complaint author himself; however, the very fact that a given complain arose usually means that there is a misrepresentation associated with the text of the given complaint. That is why the complaints are a valuable systematic source of data on deception.

To train a truth vs lie detection classified, one needs a corpora with defined ground truth. It is needed for classification tasks solving and exploring the links between implicit cues of rhetorical structure of texts and how truthful/deceptive are these texts.

The first contribution of this paper is to investigate how discourse features can be used for deception detection. The second contribution is to present the new ultimate deception dataset of bank customer complaints, it contains ground truth, is written by non-professional writers and can be used for deception detection in written texts.

The research was done for English. The paper is organized as follows. Firstly we show examples of misrepresentation in reviews and news stories, in order to highlight how it is presented in the discourse structure of texts of different genres, in both professional writing and user generated content (Sections 2, 3). Section 4 examines the existing datasets for deceptive reviews detection, it also presents briefly the main methods for deceptive texts detection, in general. In Section 5, the new dataset of customer complaints, with clear ground truth, is provided. In Section 6, we describe the deception detection methods, namely how communicative discourse trees construction and Tree Kernel learning can be applied in a system for classification of genuine/deceptive texts. Section 7 consists of first evaluation results of the classification methods, based on CDTs construction and Tree Kernel learning, on the new provided dataset, accompanied by the results on the 'gold standard' dataset of genuine/fake reviews and on the dataset from the real world. Section 8 contains conclusions.

2. Example of Misrepresentations in User-Generated Content

We provide some examples of misrepresentation in texts of different genres, in order to show how it is emphasized in the discourse structure of texts. Regarding possible misrepresentation in the user-generated content, the following example from customer complaints can be provided (1). We highlight the statement determined by the authors of this paper to be a deception in both text and its discourse tree. The statement is deceptive based on its factuality.

(1) *'I have accounts with them for almost 10 years, I hated it their customer service! Worst one ever. I don't know what's their problems, I'm not recommending their services and banking to anybody, I stopped using their credit cards already! The only reason I can't close my accounts with them, it could drop my credit score. I will not close my credit cards, but I'm not definitely using them so they can't make money from on us! I just had conversation with a supervisor from California called Steve he and his representative didn't even understand my situation, which was not common at all, basically didn't want to help me!'*

The author of this complaint does not provide a single argument backing up his claim. And the author's statement that his credit history can be negatively affected by his closing an account is a misrepresentation.

We show the text split into elementary discourse units as done by discourse parser [6]. What do we see in the discourse tree for this text? We show important (non-default) rhetorical relations in bold and highlight the verbs with the role of communicative actions which are an important addition to the rhetorical relations.

```
elaboration (LeftToRight)
 elaboration (LeftToRight)
  attribution (LeftToRight)
   TEXT:I have accounts with them for almost 10 years,
   TEXT:I hated it their customer service !
  TEXT:Worst one ever.
 elaboration (LeftToRight)
  elaboration (LeftToRight)
   explanation (LeftToRight)
    attribution (LeftToRight)
     cause (LeftToRight)
      attribution (RightToLeft)
       TEXT:I do not know
       TEXT:what is their problems,
      TEXT:I'm not recommending their services and banking to anybody,
     TEXT:I stopped using their credit cards already !
    attribution (RightToLeft)
     TEXT:The only reason I can not close my accounts with them,
     TEXT:it could drop my credit score.
   contrast (RightToLeft)
    TEXT:I will not close my credit cards,
    enablement (LeftToRight)
     TEXT:but I'm not definitely using them
     TEXT:so they can not make money from on us !
  elaboration (LeftToRight)
   TEXT:I just had conversation
   same-unit
    elaboration (LeftToRight)
     TEXT:with a supervisor from California called Steve, he and his representa-
           tive did not even understand my situation,
     TEXT:which was not common at all,
    TEXT:basically did not want to help me !
```

Figure 1. A communicative discourse tree for the user-generated text example

There is an unusual chain of rhetorical relations explanation-attribution-cause-attribution-attribution. It is a suspicious explanation pattern on its own. Unsurprisingly, the atom statement for the last attribution (which is the basis of this explanation, highlighted in Figure 1) turns out to be false.

3. Example of Misrepresentations in Professional Writing

For comparison with misrepresentation in texts written by non-professional writers, we show misrepresentation examples in news stories. In our first example, the objective of the author is to attack a claim that the Syrian government used chemical weapon in the spring of 2018 (2, **Figure 2**). An acceptable proof would be to share a certain observation, associated from the standpoint of peers, with the absence of a chemical attack. For example, if it is possible to demonstrate that the time of the alleged chemical attack coincided with the time of a very strong rain, that would be a convincing way to attack this claim. However, since no such observation was identified, the source, Russia Today, resorted to plotting a complex mental states expressing how the claim was communicated, which agents reacted which way for this communication. It is rather hard to verify most statements about the mental states of involved parties. We show the text split into EDUs as done by [6] discourse parser:

(2) *[Whatever the Douma residents,][who had first-hand experience of the shooting of the water][dousing after chemical attack video,][have to say,][their words simply do not fit into the narrative][allowed in the West,][analysts told RT.] [Footage of screaming bewildered civilians and children][being doused with water,][presumably to decontaminate them,][was a key part in convincing Western audiences] [that a chemical attack happened in Douma.] [Russia brought the people][seen in the video][to Brussels,][where they told anyone][interested in listening][that the scene was staged.] [Their testimonies, however, were swiftly branded as bizarre and underwhelming and even an obscene masquerade][staged by Russians.] [They refuse to see this as evidence,][obviously pending][what the OPCW team is going to come up with in Douma], [Middle East expert Ammar Waqqaf said in an interview with RT.] [The alleged chemical incident,][without any investigation, has already become a solid fact in the West,][which the US, Britain and France based their retaliatory strike on.]*

This article (RussiaToday 2018) does not really find counter-evidence for the claim of the chemical attack it attempts to defeat. Instead, the text says that the opponents are not interested in observing this counter-evidence. The main statement of this article is that a certain agent “disallows” a particular kind of evidence attacking the main claim, rather than providing and backing up this evidence. Instead of defeating a chemical attack claim, the article builds a complex mental states conflict between the residents, Russian agents taking them to Brussels, the West and a Middle East expert. That’s why we consider this example as misrepresentation.

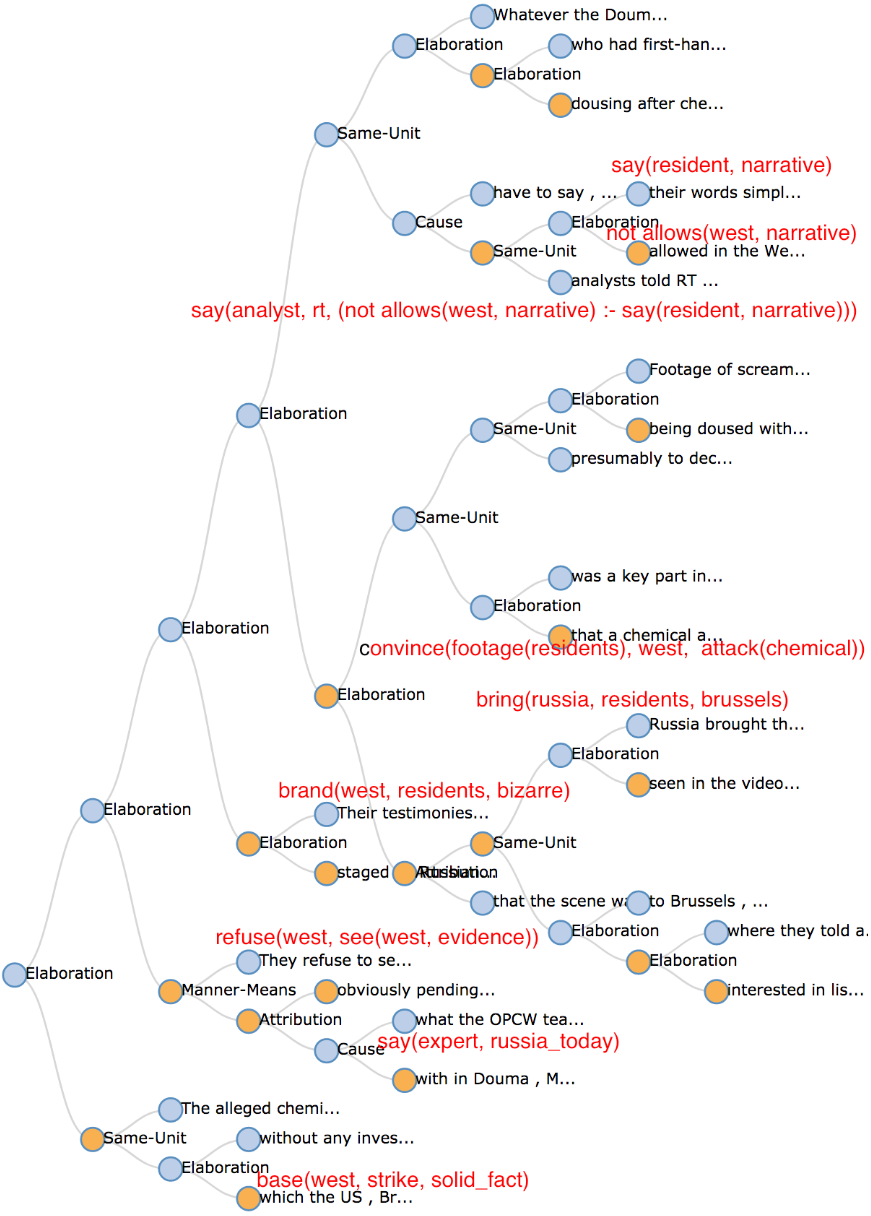

Figure 2. CDT for the chemical attack claim. An author attempts to substitute a desired valid argumentation chain by a fairly sophisticated mental states expressed by CA

Our other example of controversial news is a Trump-Russia link acquisition (BBC 2018, 3, Figure 3). For a long time it was unable to confirm the claim, so the story is repeated over and over again to maintain a reader's expectation that it would be instantiated one day. There is neither confirmation nor rejection that the dossier exists, and the goal of the author is to make the audience believe that such dossier does exist neither providing evidence nor misrepresenting events. To achieve this goal, the author can attach a number of hypothetical statements about the existing dossier to a variety of mental states to impress the reader in the authenticity and validity of the topic.

(3) *In January 2017, a secret dossier was leaked to the press. It had been compiled by a former British intelligence official and Russia expert, Christopher Steele, who had been paid to investigate Mr Trump's ties to Russia.*

The dossier alleged Moscow had compromising material on Mr Trump, including claims he was once recorded with prostitutes at a Moscow hotel during a 2013 trip for one of his Miss Universe pageants. Mr Trump emphatically denies this.

The file purported to show financial and personal links between Mr Trump, his advisers and Moscow. It also suggested the Kremlin had cultivated Mr Trump for years before he ran for president.

Mr Trump dismissed the dossier, arguing its contents were based largely on unnamed sources. It was later reported that Mr Steele's report was funded as opposition research by the Clinton campaign and Democratic National Committee.

Fusion GPS, the Washington-based firm that was hired to commission the dossier, had previously been paid via a conservative website to dig up dirt on Mr Trump.

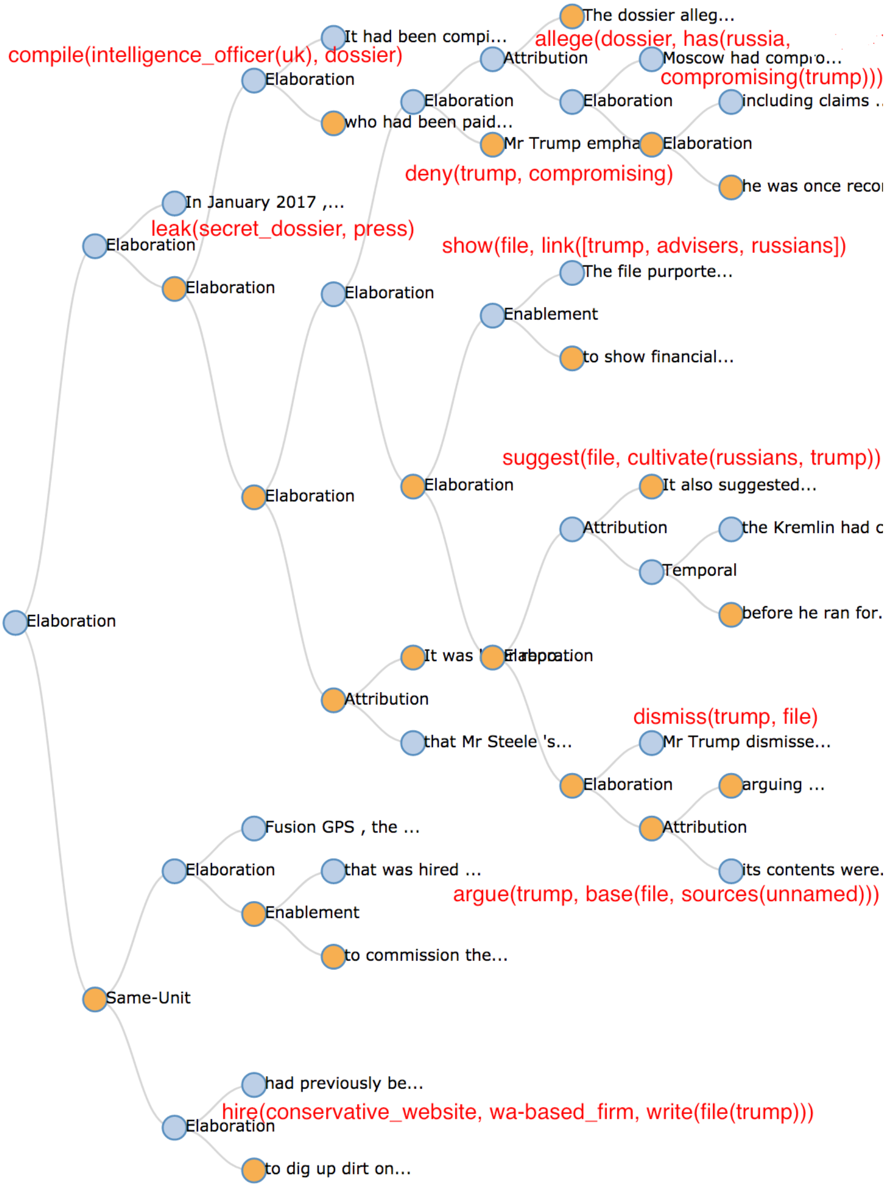

Figure 3. CDT for an attempt to prove something where an evidence is absent so the facts are “wrapped” into complex mental states as expressed by communicative actions

4. Background and Related Work on Deception Datasets

As customer complaints are a subgenre of reviews, we pay the main attention to the existing truthful/deceptive reviews datasets. Deceptive product reviews can be referred to as deceptive opinion spam: fictitious opinions that have been deliberately written to sound authentic, in order to deceive the reader [11]. Spammers write fake reviews to promote or demote target products. They are deliberately written in order to sound authentic, and it is difficult to recognize them manually: human average accuracy is merely 57.3% [11].

Automated deception detection for reviews faces the lack of 'gold standard' corpora with verified examples of deceptive uses of language. Besides this, intentionally written (e.g. by crowdsourcing) texts are distinct from genuinely produced texts. Hence, such artificial texts classified as deceptive by human annotators are not necessarily totally deceptive.

The release of two 'gold standard' datasets (available at http://myleott.com/) allowed for applying supervised learning methods, taking stylistic, syntactic and lexical features into consideration [12], [11], [13], [14]. Hotels reviews were chosen for the datasets, because it was suggested that deception rates among travel reviews is reasonably small. The latter dataset includes, among other reviews, crowdsourced generation of deceptive reviews. It contains 400 truthful positive reviews from TripAdvisor; 400 deceptive positive reviews from Mechanical Turk; 400 truthful negative reviews from reviews websites; 400 deceptive negative reviews from Mechanical Turk.

Later researchers tried to overcome the lack of large realistic datasets on different topics and domains. For example, Yao et al. [15] apply a data collection method based on social network analysis to quickly identify deceptive and truthful online reviews from Amazon. The dataset contains more than 10,000 deceptive reviews in diverse product domains. The problem of the mentioned 'gold standard' datasets is that the fake reviews were not taken from genuinely written ordinary reviews and manually classified as fake. Instead, whey were written on demand by the Amazon Mechanical Turk workers, hence they are not indicative of deception [16]. However, they are accepted as 'gold standard' datasets for this research field. Rules used in [12] to create ground truth datasets were used in later projects, such as in [17].

The real-life Amazon dataset [18] contains reviews from Amazon.com (crawled in 2006) which is large and covers a very wide range of products. It was used, for example, in Sun et al. [19], namely, three domains: Consumer Electronics, Software, and Sports. The metadata in this dataset provides only helpfulness votes of the reviews.

In cases where there was no certain knowledge of the ground truth, different ways to collect reviews corpora, relying on other features, were used. For example, in [14] the DeRev corpus of books reviews, originally posted on Amazon, was collected using definite pre-defined deception clues, Book reviews in the corpus are marked as clearly fake, possibly fake, and possibly genuine. The corpus is constituted by 6,819 instances whose 236 were labeled with the higher degree of confidence and are considered as the 'gold standard'.

In [20], two publicly available Yelp datasets were presented. They are labeled with respect to the Yelps classification in recommended and not recommended reviews. Mukherjee et al. [21] found that the Yelp spam filter primarily relies on linguistic, behavioral, and social networking features. Classification provided by Yelp has been also used in many previous works before as a ground truth, where recommended reviews correspond to genuine reviews, and not recommended reviews correspond to fake ones, so these labels can be trusted. The Yelp NYC dataset contains reviews of restaurants located in New York City (359,052 reviews; 10.27% are fake); the Zip dataset is larger, since it contains businesses located in contiguous regions of the U.S. (608,598 reviews; 13.22% are fake).

Big Amazon dataset is annotated with compliant/non-compliant labels. It has many different topics: from electronics and books to office products (https://s3.amazonaws.com/amazon-reviews-pds/readme.html). It contains labels about star rating, helpful vote, total votes, verified purchase. That could be used for making decisions.

Hence, the existing recent datasets rely on external factors provided by their source, such as review's rating, number of votes, social networking features of review's author, metadata features etc. They are not annotated manually. So, despite the presence of different corpora, lack of corpora with exact ground truth can be understood as a bottleneck in deception detection of online reviews and similar text genres.

For fake reviews detection, language features and behavioral features are usually used, as in [22], [23]. The impact of different language features on deception detection, in general, was studied in [24], [25]. In recent years, big amounts of news stories with misinformation caused by political reasons [26] led to the specific attention to fake news detection studies for English. Several new datasets were proposed, as in [27], [28], [29]. In [30], the combined approach, based on language features, was suggested: there are linguistic (n-gram), credibility-related (capitalization, punctuation, pronoun use, sentiment polarity), and semantic (embeddings and DBPedia data) features. Close approach based on a set of various language features was suggested in [31] (ngrams, punctuation, psycholinguistic features, readability, syntax) and [32] (stylistic, complexity, psychological features). Deep learning approaches were used in [28], [33]. Source and web page features were added in [34], [35]. As to language features, unlike lexical, syntactic and semantic features, discourse features are less used due to the complexity of the approach. Despite this, automated fake news detection, based on simple discourse features, was studied in [36] and is included in the proposed methods for deception detection in written texts. Hence, we decided to examine if more complex discourse features could be useful for automated deception detection in case of reviews and complaints.

5. Description of the Training Dataset

We introduce the ultimate deception dataset. It contains customer complaints—emotionally charged texts which are very similar to reviews and include descriptions of problems they experienced with certain businesses. Raw complaints in English were collected from PlanetFeedback.com for a number of banks submitted in 2006–2010. The dataset consists of 2,746 complaints totally. 400 complaints were manually tagged with respect to the parameters related to argumentation and validity of text: perceived complaint validity; argumentation validity; presence of specific argumentation patterns; detectable misrepresentation. Here, validity of information is connected with validity of arguments. The dataset contains texts with direct truth confirmation based on manual annotation. It contains authentic data: both truthful and deceptive reviews were taken from spontaneously written customers' texts. Among the annotated 400 complaints, 163 contain a deception.

This dataset includes more emotionally-charged complaints in comparison with other argument mining datasets, such as [37], [38], [39]. For a given topic such as insufficient funds fee, this dataset provides many distinct ways of argumentation that this fee is unfair. Authors attempt to provide as strong argumentation as possible to back up their claims and strengthen their case.

If a complaint is not truthful, it is usually invalid: either a customer complains out of a bad mood or wants to get a compensation. However, if the complaint is truthful it can easily be invalid, especially when arguments are flawed. When an untruthful complaint has valid argumentation patterns, it is hard for an annotator to properly assign it as valid or invalid, without the guidelines. So, according to the guidelines for the manual tagging of the dataset, a complaint was considered as valid if a judge believed that the main complaint claim is truthful under the assumption that a complainant is making truthful statement. Valid complaint needs to include proper discourse and acceptable argumentation patterns. Following this approach, a complaint is marked as truthful if a judge cannot defeat it, using commonsense knowledge, available factual knowledge about a domain or implicit, indirect cues. Inconsistencies detected by a judge also indicate that the complaint author is deceiving. Mentioning multiple unusual, very rarely occurring claims also indicate that the complaint author is deceiving. The judge does not have to be able to prove that the complainant is lying: judge's intuition is sufficient to tag a complaint as untruthful. We suggest that one can provide a valid argumentation and also provide a false statement in a single sentence: 'Rule is like this <correct rule> and I followed it, making <false statement>. Conversely, one can be truthful but provide an invalid argumentation pattern "I set this account for direct deposit and sent a check out of it <truthful statement>, as my HR manager suggested <should not have followed advice from not a specialist in banking>. Therefore validity (of argumentation patterns) and truthfulness are correlated.

Initial set of 400 complaints was tagged by the authors of the paper as experts. After that, three annotators worked with this dataset, having a set of definitions and applying them. Then precision and recall were measured by matching the tags done by the authors as the 'gold standard', after that the set of definitions was edited and elaborated. In the further work, the Krippendorff's alpha measure (for three annotators) was applied as inter-annotator agreement measurement, and it exceeds 80%. Complaints reveal shady practice of banks during the financial crisis of 2007—for instance, manipulating an order of transactions to charge a highest possible amount of non-sufficient fund fees. As it is possible to know, retrospectively and based on facts, the established ground truth, we suggest that the annotators can find out, with high confidence, what information in texts is deceptive. So the dataset would provide ground truth.

The rest complaints were auto-tagged based on the model trained on this 400 set. Then they have been partially manually evaluated. The accuracy of auto tagging exceeds 75%, so these labeled complaints can be also used for the classifiers training.

Customer complaints can be considered as a subgenre of reviews in general, but despite this complaints have much more significance for well-being of customers in comparison with customer reviews. Furthermore, customer complaints have much more significance for well-being of customers in comparison with customer reviews. Therefore, tagged customer complaints have much more importance associated with truth/deception than customer reviews. Since reviews are associated with opinions which can be random and complaints with customers doing their best to achieve their goals, both the truth and a lie is much more meaningful and serious in comparison with review datasets.

Complaints usually have a simple motivational structure; they are written with an obvious goal. Most complainants are faced with a strong deviation between what they expected from a service, what they received and how it was communicated. Most complaint authors report incompetence, flawed policies, ignorance, indifference to customer needs from the customer service personnel. The authors are frequently exhausted communicative means available to them, confused, seeking recommendation from other users and advising others on avoiding particular financial service. The focus of a complaint is a proof that the proponent is right and the opponent is wrong, as well as resolution proposal and a desired outcome.

6. Detecting Deception via Communicative Discourse Trees

In the Rhetorical Structure Theory [1], [7], discourse is understood as a hierarchical system of discourse units of different size, where smaller discourse units can be successively incorporated into larger ones. Discourse unites can be combined into a higher unit in case there is a rhetorical (discourse) relation of a certain type between them, e.g. *Concession*, *Elaboration*. One of the discourse units is the nucleus (more important), while the other is a satellite (contains the additional information). An elementary discourse unit (EDU) usually corresponds to a clause.

Two RST parsers constructing discourse tree (DT) from paragraphs of text are available at the moment. We used the tool provided by [6], [8]. After that, we build CDTs involving VerbNet.

Argumentation analysis needs a systematic approach to learn associated discourse structures. The features of CDTs could be represented in a numerical space so that argumentation detection can be conducted; however, structural information on DTs would not be leveraged. Also, features of argumentation can potentially be measured in terms of maximal common sub-DTs, but such nearest neighbor learning is computationally intensive and too sensitive to errors in DT construction. Therefore, a CDT-kernel learning approach is selected which applies a support vector machine (SVM) learning to the feature space of all sub-CDTs of the CDT for a given text where an argument is being detected.

Tree Kernel (TK) learning for strings, parse trees and parse thickets is a well-established research area nowadays. The CD-TK counts the number of common subtrees as the discourse similarity measure between two DTs. In this study, we extend the TK definition for the CDT, augmenting DT kernel by the information on CAs. TK-based approaches are not very sensitive to errors in parsing (syntactic and rhetorical) because erroneous sub-trees are mostly random and will unlikely be common among different elements of a training set.

A CDT can be represented by a vector V of integer counts of each sub-tree type (without taking into account its ancestors):

$$V(T) = (\#\ of\ subtrees\ of\ type\ 1, \ldots, \#\ of\ subtrees\ of\ type\ I, \ldots, \#\ of\ subtrees\ of\ type\ n).$$

Given two tree segments CDT_1 and CDT_2, the tree kernel function is defined:

$$K\ (CDT_1, CDT_2) = < V(CDT_1), V\ (CDT_2) > = \Sigma_i V(CDT_1)[i], V(\mathrm{CDT}_1)[i] = \Sigma n_1 \Sigma n_2\ \Sigma_i I_i(n_1) \times I_i(n_2),$$

where $n_1 \in N_1$, $n_2 \in N_2$ and N_1 and N_2 are the sets of all nodes in CDT_1 and CDT_2, respectively; $I_i(n)$ is the indicator function:

$$I_i(n) = \{1\ if\ a\ subtree\ of\ type\ i\ occurs\ with\ a\ root\ at\ a\ node;\ 0\ otherwise\}.$$

Further details for using TK for paragraph-level and discourse analysis are available in [9].

Only the arcs of the same type of rhetorical relations (presentation relation, such as antithesis, subject matter relation, such as condition, and multinuclear relation, such as List) can be matched when computing common sub-trees. We use N for a nucleus or situations presented by this nucleus, and S for a satellite or situations presented by this satellite. Situations are propositions, completed actions or actions in progress, and communicative actions and states (including beliefs, desires, approve, explain, reconcile and others). Hence we have the following expression for RST-based generalization '^' for two texts $text_1$ and $text_2$:

$$text_1 \wedge text_2 = \cup_{i,j}(rstRelation_{1i}, (\ldots, \ldots) \wedge rstRelation_{2j}(\ldots, \ldots)),$$

where $i \in (RST\ relations$ in $text_1)$, $j \in (RST\ relations$ in $text_2)$. Further, for a pair of RST relations their generalization looks as follows:

$$rstRelation_1(N_1, S_1) \wedge rstRelation_2\ (N_2, S_2) = (rstRelation_1 \wedge rstRelation_2)\ (N_1 \wedge N_2, S_1 \wedge S_2).$$

We define CA as a function of the form verb (agent, subject, cause), where verb characterizes some type of interaction between involved agents (e.g., explain, confirm, remind, disagree, deny, etc.), subject refers to the information transmitted or object described, and cause refers to the motivation or explanation for the subject. To handle meaning of words expressing the subjects of CAs, we apply word2vec models [10].

For EDUs as labels for terminal nodes only the phrase structure is retained. The terminal nodes are labeled with the sequence of phrase types instead of parse tree fragments.

We combined Stanford NLP parsing, coreference resolution tool, entity extraction, DT construction (discourse parser), VerbNet and Tree Kernel builder into one system.

The system is available at https://github.com/bgalitsky/relevance-based-on-parse-trees with the more detailed description. It can be used for similar tasks.

For EDUs as labels for terminal nodes only the phrase structure is retained: we suppose to label the terminal nodes with the sequence of phrase types instead of parse tree fragments. For the evaluation purpose Tree Kernel builder tool [5] was used. These discourse trees features are given to the classifiers.

7. Evaluation Results

We first train the deception detection model on our ultimate deception dataset. For the initial and automatically derived datasets, we show the accuracies of training (grayed) row and testing, averaging through 5x cross-validation. For the bottom three datasets, we only tested the obtained model. For genuine reviews, 380 cases of deception were detected which were false positives, assuming that review writers do not lie (**Table 1**).

Table 1: Datasets, evaluation settings and recognition accuracies for deception detection

Dataset	Deception	No deception	Precision	Recall	F1 score
Manually tagged complaints	163	237	91	85	88
			83	81	82
Automatically tagged based on initial classifier	1,132	1,615	78	75	76
			69	71	70
Genuine reviews	580	3,420	83	100	91
Fake reviews	414	286	100	59	74
Enron	27	10,000	85	0.1 (estimated)	0.2

We explored whether fake opinionated text have a similar rhetorical structure to text with deception, and genuine reviews have similar rhetoric structure to texts without deception. We took the ‘gold standard’ reviews dataset: fake reviews and genuine reviews [11], [12] (**Table 1**).

In [11], [12] authors addressed the problem of detection of opinion spam: obvious instances that are easily identified by a human reader, including advertisements, questions, and other irrelevant or non-opinionated texts. The authors investigated a more implicit type of opinion spam such as deceptive opinion spam, ones that have been deliberately written to sound authentic, in order to deceive the reader. Fake reviews were written by Amazon Mechanical Turk workers. The instructions asked the workers to assume that they are employed by a hotel’s marketing department, and to pretend that they are asked to write a fake review (as if they were a customer) to be posted on a travel review website; additionally, the review needs to sound realistic and portray the hotel in a positive light. A request for negative reviews was done analogously.

Although our SVM TK system did not achieve [11], [12] performance of 90% on their data, the task of detection of fake review texts as the ones including deception was performed at 74% accuracy by the classifier.

We suggest that the system could be applied to different text genres (written by non-professional writers), so it could be the universal text classification system for deception, the same which extracts arguments and assesses sentiments polarity. Hence, we run the following evaluation experiment in order to start checking this point.

To assess the deception detection in a real world deception-neutral environment, we ran our detector again the business communication dataset of Enron [40], using it as the evaluation dataset. This dataset represents neither user-generated content since this is work-related correspondence, not professional writing since the email authors are employees of an organization with various roles. Naturally, deception is concealed, and we do not know what was actually happening in the company and among its business partners. However, a small number of interesting email have been discovered which have a peculiar logical structure and might well be a misrepresentation. Annotators looked at them manually to understand if they were similar to misrepresentation, although we did not have ground truth here. They could not be sure if reviews with tricky patterns similar to misrepresentation were really misrepresentation, but the detector could identify possible reviews with misrepresentation, that were also identified as the 'suspicious' ones (containing possible misrepresentation) by human annotators. Precision turned out to be high and recall extremely low since only a small fraction of deception emails has been discovered. The resultant 0.2% F-score is not an indication of recognition accuracy but instead of our available estimate of the classes in the Enron dataset.

We do not know the actual proportion of emails with misrepresentation in Enron dataset but all detected cases are important since a misrepresentation is uncovered. Recall is not as important for this task as precision: we want to avoid false positives: once an email is classified as the one with deception we would expect to manually confirm it.

We now zoom into the deception detection methodology for the most adequate case, the set of 2,747 automatically tagged complaints (Table 2).

Table 2: Classification accuracy for the baseline and the approach being proposed for deception detection

Method	Precision	Recall	F1 score
Keyword-based	56	53	54
Naïve Bayes	61	63	62
SVM-TK over parse trees and DTs	67	69	68
SVM-TK over parse trees and DTs labeled with CAs	69	71	70

One can see that keyword-based and Naive Bayes classifier perform slightly better than random, since deception manifests itself at the discourse level, not the syntactic one. Then we observe that proceeding to machine learning of DTs delivers 8% gain in classification accuracy.

A deep learning approach could be potentially applied to our structured representation. However, based on our experience with discourse-level data that the amount and quality of data contributes significantly more to the overall accuracy of a classifier, we believe experiments with the same data but different machine learning framework would be redundant.

8. Conclusions

An extensive corpus of literature on RST parsers does not address the issue of how the resultant DT will be employed in practical NLP systems. RST parsers are mostly evaluated with respect to agreement with the test set annotated by humans rather than its expressiveness of the features of interest. In this work we focused on interpretation of DT and explored ways to represent them in a form indicative of a conflict rather than neutral enumeration of facts.

In several previous papers about SVM TK and discourse, it was observed that using SVM TK, one can differentiate between a broad range of text styles, genres and abstract types. These classes of texts are important for a broad spectrum of applications of recommendation and security systems, from finance to data loss prevention domains. Each text style and genre has its inherent rhetorical structure which is leveraged and automatically learned. Since the correlation between text style and text vocabulary is rather low, traditional classification approaches which only take into account keyword statistics information could lack the accuracy in the complex cases.

We showed that deception detection methodology based on rhetorical structure of texts, being applied to various text genres—news texts, online reviews, customer complaints, business communication texts—seems promising and needs to be investigated further. Next steps for proving the hypothesis of a deception being visible from a text's discourse structure should be done. Here, further experiments based on the presented ultimate deception dataset of bank customer complaints should be held. This dataset is in the initial stage now and is still being developed. In the future studies, the whole complaint dataset should be manually annotated. The recognition method will be applied to a bigger annotated dataset part. Results obtained on this dataset should be also compared with other results obtained on 'gold standard' datasets. For a bigger dataset training, we could also apply deep learning models. We will also focus on more experiments for precision improvements, as reducing the number of false positives is mostly important for deception detection task. We also plan to run further experiments on different text genres, to check if the universal text classification system for deception, based on discourse features, could be universal. Both truthfulness and validity are recognized reasonably well which is a value for Customer Relation Management systems and could be useful in different NLP tasks that are based on online reviews analysis.

Acknowledgements

This paper is partially supported by Russian Foundation for Basic Research (project No. 17-29-07033).

References

1. *William C. Mann and Sandra A. Thompson* (1988), Rhetorical structure theory: Toward a functional theory of text organization. Text-Interdisciplinary Journal for the Study of Discourse, 8(3), pp. 243–281.
2. *Boris Galitsky, Dmitry Ilvovsky, and Dina Pisarevskaya* (2018), Argumentation in text: Discourse structure matters, CICLing 2018 (unpublished).
3. *Grasso Floriana* (2003), Characterizing Rhetoric Argumentation, PhD Thesis HERIOT-WATT UNIVERSITY.
4. *B. Galitsky, D. Ilvovsky, and T. Makhalova* (2019), Enabling a Bot with Understanding Argumentation and Providing Arguments, Developing Enterprise Chatbots, Springer—Cham, Switzerland.
5. *B. Galitsky. Rhetorical Agreement: Maintaining Cohesive Conversations* (2019), Developing Enterprise Chatbots, Springer—Cham, Switzerland.
6. *S. Joty, G. Carenini, R. T. Ng, and Y. Mehdad* (2013), Combining intra-and multisentential rhetorical parsing for document-level discourse analysis, ACL (1), pp. 486–496.
7. *M. Taboada, W. C. Mann* (2006), Applications of rhetorical structure theory, Discourse Studies, 8(4), pp. 567–588.
8. *M. Surdeanu, T. Hicks, and M. A. Valenzuela-Escarcega* (2015), Two Practical Rhetorical Structure Theory Parsers, NAACL HLT.
9. *Galitsky B.* (2017), Improving relevance in a content pipeline via syntactic generalization, Engineering Applications of Artificial Intelligence 58, pp. 1-26..
10. *Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, Armand Joulin* (2017), Advances in Pre-Training Distributed Word Representations, LREC 2018.
11. *Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock* (2011), Finding deceptive opinion spam by any stretch of the imagination, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, volume 1, pp. 309–319.
12. *Myle Ott, Claire Cardie, and Jeffrey T. Hancock* (2013), Negative deceptive opinion spam, NAACLHLT 2013, Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 497–501.
13. *S. Feng, R. Banerjee, and Y. Choi* (2012), Syntactic stylometry for deception detection, ACL 12, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pp. 171–175.
14. *Tommaso Fornaciari and Massimo Poesio* (2014), Identifying fake amazon reviews as learning from crowds, Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pp. 279–287.
15. *Wenlin Yao, Zeyu Dai, Ruihong Huang, and James Caverlee* (2017), Online deception detection refueled by real world data collection, Proceedings of Recent Advances in Natural Language Processing, pp. 793–802.
16. *A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance* (2013), Fake review detection: classification and analysis of real and pseudo reviews. tech. rep. uic-cs-2013-03. University of Illinois at Chicago.

17. *Z. Hai, P. Zhao, P. Cheng, P. Yang, X.-L. Li, and G. Li* (2016), Deceptive review spam detection via exploiting task relatedness and unlabeled data, Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1817–1826.
18. *Nitin Jindal and Bing Liu* (2008), Opinion spam and analysis, Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM '08). ACM, New York, NY, USA, pp. 219–230.
19. *Chengai Sun, Qiaolin Du, and Gang Tian* (2016), Exploiting product related review features for fake review detection, Mathematical Problems in Engineering.
20. *S. Rayana and L. Akoglu* (2015), Collective opinion spam detection: Bridging review networks and metadata, Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 985–994.
21. *A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance* (2013), What yelp fake review filter might be doing?, Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media.
22. *J. Fontanarava, G. Pasi, and M. Viviani* (2017), Feature Analysis for Fake Review Detection through Supervised Classification, 2017 International Conference on Data Science and Advanced Analytics, pp. 658–666.
23. *S. Mukherjee* (2017), Probabilistic Graphical Models for Credibility Analysis in Evolving Online Communities. Doctor Thesis.
24. *E. Fitzpatrick, J. Bachenko, and T. Fornaciari* (2015), Automatic Detection of Verbal Deception. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers.
25. *P. Rosso and L. Cagnina* (2017), Deception Detection and Opinion Spam, A Practical Guide to Sentiment Analysis, Cambria, E., Das, D., Bandyopadhyay, S., Feraco, A. (Eds.), Socio-Affective Computing, vol. 5, Springer-Verlag, pp. 155–171.
26. *H. Allcott and M. Gentzkow* (2017), Social Media and Fake News in the 2016 Election., Journal of Economic Perspectives, Vol. 31–2, pp. 211–236.
27. *Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi* (2017), Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking, Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2931–2937.
28. *William Yang Wang* (2017), "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 422–426.
29. *Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu* (2018), FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media, available at: https://arxiv.org/abs/1809.01286.
30. *Momchil Hardalov, Ivan Koychev and Preslav Nakov* (2016), In Search of Credible News, Artificial Intelligence: Methodology, Systems, and Applications. AIMSA 2016. Lecture Notes in Computer Science, vol 9883. Springer, Cham, pp. 172–180.
31. *Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea* (2018), Automatic Detection of Fake News, Proceedings of the 27th International Conference on Computational Linguistics, pp. 3391–3401.

32. *Benjamin D. Horne, and Sibel Adali* (2017), This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News, available at: https://arxiv.org/abs/1703.09398.
33. *Natali Ruchansky, Sungyong Seo, and Yan Liu* (2017), CSI: A Hybrid Deep Model for Fake News Detection, CIKM'17 Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 797–806.
34. *Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov* (2018), Predicting Factuality of Reporting and Bias of News Media Sources, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3528–3539.
35. *Benjamin D. Horne, William Dron, Sara Khedr, Sibel Adalı* (2018), Assessing the News Landscape: A Multi-Module Toolkit for Evaluating the Credibility of News, WWW 2018, April 23–27, 2018, Lyon, France, pp. 235–238.
36. *Rubin, V. L. Conroy, N. J. and Chen Y. C.* (2015), Towards News Verification: Deception Detection Methods for News Discourse, Proceedings of the Hawaii International Conference on System Sciences (HICSS48) Symposium on Rapid Screening Technologies, Deception Detection and Credibility Assessment Symposium, January 5–8, 11 pages.
37. *Christian Stab and Iryna Gurevych* (2017), Recognizing insufficiently supported arguments in argumentative essays, Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017), pp. 980–990.
38. *Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn A. Walker* (2016), Internet argument corpus 2.0: An sql schema for dialogic social media and the corpora to go with it, Language Resources and Evaluation Conference.
39. *Shereen Oraby, Lena Reed, Ryan Compton, Ellen Riloff, Marilyn Walker, and Steve Whittaker* (2015), And thats a fact: Distinguishing factual and emotional argumentation in online dialogue, NAACL HLT 2015 2nd Workshop on Argumentation Mining.
40. *Cohen W. W.* (2019), Enron Email Dataset, available at: https://www.cs.cmu.edu/~./enron/.