Information Processing and Management 56 (2019) 102102

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Automatic classification of complaint letters according to service provider categories

Yaakov HaCohen-Kerner[a,*], Rakefet Dilmon[b], Maor Hone[a], Matanya Aharon Ben-Basan[b]

[a] Dept. of Computer Science, Jerusalem College of Technology, 21 Havaad Haleumi St., P.O.B. 16031, 9116001 Jerusalem, Israel
[b] Dept. of Hebrew and Semitic Languages, Bar-Ilan University, 5290002 Ramat-Gan, Israel

ARTICLE INFO

Keywords:
Bag of words
Complaint letters
Semantic fields
Service providers
Supervised machine learning
Text classification

ABSTRACT

In the technological age, the phenomenon of complaint letters published on the Internet is increasing. Therefore, it is important to automatically classify complaint letters according to various criteria, such as company categories. In this research, we investigated the automatic text classification of complaint letters written in Hebrew that were sent to various companies from a wide variety of categories. The classification was performed according to company categories such as insurance, cellular communication, and rental cars. We conducted an extensive set of classification experiments of complaint letters to seven/six/five/four company categories. The classification experiments were performed using various sets of word unigrams, four machine learning methods, two feature filtering methods, and parameter tuning. The classification results are relatively high for all six measures: accuracy, precision, recall, F1, PRC-area, and ROC-area. The best accuracy results for seven, six, five, and four categories are 84.5%, 88.4%, 91.4%, and 93.8%, respectively. An analysis of the most frequently occurring words in the complaints about almost all categories revealed that the most significant issues were related to poor service and delayed delivery. An interesting result shows that only in the domain of hospitals was the subject of the domain itself (i.e., the patient, the medical treatment, the place of the treatment, and the medical staff) the most important issue. Another interesting finding is that the issue of "price" was of little or no importance to the complainants. These findings suggest that in their preoccupation with their bottom line of profitability, many service providers are blind to how paramount good service and timely delivery (and, in the case of hospitals, the domain itself) are to their clientele.

1. Introduction

Automatic text classification (TC) is a highly researched domain. According to Google Scholar,[1] about 11,000 academic papers include the terms "text" along with "classification" or "categorization" in their titles. An example of an interesting text classification task is the automatic classification of complaint letters.

In recent years, two parallel processes have been taking place. On the one hand, customers are becoming more demanding and willing to complain about the goods and services they purchase (Mahayudin, Haron & Yin-Fah, 2010). On the other hand, more and

* Corresponding author.
E-mail addresses: kerner@jct.ac.il (Y. HaCohen-Kerner), ak2@bezeqint.net (R. Dilmon).

[1] Results of Google Scholar search papers include the terms "text" along with "classification" or "categorization" in their titles on 13/AUG/19

https://doi.org/10.1016/j.ipm.2019.102102
Received 14 February 2018; Received in revised form 13 August 2019; Accepted 13 August 2019
Available online 29 August 2019
0306-4573/

more companies are adopting social media as part of their communication strategies (Avery, Steenburgh, Deighton & Caravella, 2012). Some of the companies integrate social media in their complaint handling process, e.g., a German telecommunication services provider (Rossmann, Wilke & Stei, 2017). The customers of this company, immediately after a service experience, received an invitation to take part in a service survey by direct message (Twitter) or by direct mail (Facebook). The findings of Rossmann et al. (2017) illustrate that it is not always in the company's best interest to use social media channels. Though use of social media systems is expected to reduce the customer effort and improve the quality of solutions, the marginal impact on satisfaction and subsequently on behavioral intentions is higher using traditional media.

Pinto and Mansfield (2012) investigated the complaining behavior of online Generation Y users in general and their use of Facebook as a complaint channel in particular. Significant differences were found between males and females in three categories: mobile phone shops, clothing stores, and hair stylists. In each of these categories, women used Facebook much more than men to complain about the service provider or the product.

Social media (e.g., blogs, forums, tweets, and websites) provide a platform for companies to be in contact with their customers and to learn from them (Jin, Yan, Yu & Li, 2013). According to Tripp and Gregoire (2011, p. 37), "companies need to understand and manage the rising threat of online public complaining." Both the academic and the practitioner worlds recognized customer satisfaction and consumer complaining behavior as an important phenomenon that greatly influences the success of an organization (Pinto & Mansfield, 2012). Moreover, several studies identified successful complaint resolution as a competitive advantage (Fox, 2008; Tax, Brown & Chandrashekaran, 1998).

The problem is that there are hardly any studies, articles, or systems dealing with automatic classification of complaint letters. This need for automatic classification of complaint letters is very important because the manual identification of complaints is inefficient and time-consuming, due to the significant increase in complaints regarding a wide range of topics with sparse distribution. Maia, Carvalho, Ladeira, Rocha and Mendes (2014), for example, reported that there are more than 6000 complaints in the database servers of Brazil's Office of the Comptroller General (Controladoria Geral da União – CGU) that are waiting to be screened and analyzed. On average, Brazilian citizens submit 32 new complaints each day, while CGU can analyze only around 20 per day. At this rate, in a year CGU will have about 10,400 complaints still waiting to be analyzed, an increase of 73% (relative to 6000 complaints that are waiting to be analyzed at the beginning of the year). Clearly, without a major change in its screening process, CGU will not be able to analyze all submitted complaints within a reasonable period of time.

In this study, we investigate the issue of automatic text classification of complaint letters written in Hebrew (a domain that has not been studied to this day). In particular, we investigate the classification of complaint letters that were sent to various companies in a wide variety of categories.

The general research objectives of this study are: (1) generate a classifier for category specific online complaints in Hebrew, and (2) identify the characteristics of the complaints for each category. While classifiers for complaints have been partially addressed, no work on Hebrew complaints has been done., The companies can use the characteristics of the complaints to identify the features of the complaint letters in their category in general, and specifically against their company, in order to improve their future service.

Our approach is as follows. We built various variants of classification models that were based on an extensive set of classification experiments on complaint letters in seven/six/five/four company categories. The classification experiments were performed with various sets of word unigrams, four machine learning (ML) methods, two feature filtering methods, and parameter tuning, combined with extensive data analysis. In order to facilitate analysis of the characteristics of the complaint letters in each category, we extracted representative semantic fields for each category (a "'semantic field"' is a group of words with a common core of meaning, e.g., service provider-customer relations, price, time, location, and subject domain). The semantic fields enable a deeper analysis of the characteristics of the complaint letters for each category. For the public, this analysis renders the companies' attitudes toward customers transparent; for the companies, it aids focusing on the most important fields and improving their service.

The main contributions of this study are (1) the first supervised dataset, to the best of our knowledge, of complaint letters written in Hebrew related to companies from various categories, e.g., insurance, cellular communication, and rental cars; (2) a successful model capable of classifying complaint letters according to company categories; and (3) findings that indicate which semantic fields and which words in the various categories matter the most to the complainants.

The structure of the article is as follows. Section 2 introduces text classification and previous studies in the domain of application of TC to complaint letters. Section 3 presents the methodology. Section 4 presents the constructed corpus of complaint letters. Section 5 introduces the model and many experimental results. Section 6 presents a semantic analysis of selected results. Finally, Section 7 summarizes, and offers some suggestions for future research.

2. Related work

In this section, after defining TC, we describe a traditional model for it. Then we present several TC studies, feature filtering methods, and recent TC innovations. Finally, we describe several recent studies in the field of application of TC to complaint letters.

2.1. Text classification

TC is the task of automatically assigning documents to a fixed number of categories (Joachims, 1998). In principle, each document can be classified into one or more categories, with the typical case being to classify each document into only one category. Classification algorithms typically use a supervised ML method, or a combination of several ML methods (Sebastiani, 2002). TC is an important component in many research domains, e.g., information extraction, information retrieval, question answering, sentiment

analysis, text indexing, text mining, and word sense disambiguation (Knight, 1999; Lee & Dernoncourt, 2016; Patra & Singh, 2013).

The traditional model for domain-based TC is based on a bag-of-words (BOW) representation, which associates a text with a vector indicating the number of occurrences of each chosen word in the training corpus. Various ML methods such as Naive Bayes (NB) (McCallum & Nigam, 1998), Support Vector Machines (SVM) (Cortes & Vapnik, 1995), and Maximum Entropy (ME) (Mayer, 1960) have been reported to use BOW features to achieve accuracies of 90% and above for particular categories (Joachims, 1998). The BOW features are regarded as simple (i.e., simple definition and easily extracted) but efficient, and even very efficient for certain TC tasks (Shahana & Omman, 2015; Yamasaki et al., 2015). The linear ordering of the words within the context is ignored (i.e., the BOW representation is essentially independent of the sequence of words in the collection). Word n-grams with a length of $n > 1$ are ignored in the BOW representation.

Examples of TC studies according to categories (e.g., disciplines, domains, and topics) using various types of BOWs, and general features such as word n-grams and character n-grams are proposed by Fürnkranz (1998), Martins and Silva (2005), HaCohen-Kerner, Mughaz, Beck, & Yehudai (2008), HaCohen-Kerner, Dilmon, Friedlich and Cohen (2016), Liparas, HaCohen-Kerner, Moumtzidou, Vrochidis and Kompatsiaris (2014), and HaCohen-kerner, Ido and Ya'akobov (2017).

Mladenic and Grobelnik (1998) and Fürnkranz (1998) found that the addition of word bigrams and tri-grams to the BOW representation improves the performance of TC. However, sequences of length $n > 3$ were found to be useless. The disadvantage of word n-gram features is that they increase the dimensionality of the problem because of the high number of combinations of word n-grams. The linear ordering of the word n-grams within the context is ignored.

Kanaris, Kanaris, Houvardas and Stamatatos (2007) showed that character n-grams discriminate between spam and legitimate email messages better than word n-grams, even though character n-grams increase the dimensionality of the problem. The authors found that the most important property of the character n-gram method is that it avoids the use of lemmatizers, tokenizers, and other language-dependent tools. Kanaris and Stamatatos (2009) showed that character n-grams are better genre discriminators than word unigrams for automatic detection of webpage genres. Additional advantages of character n-grams are that they are language-independent and easily extracted. The main disadvantage of the character n-gram features is that they considerably increase the dimensionality of the problem.

An important pre-processing task in TC is the application of feature filtering methods. Feature filtering aims to remove either irrelevant or redundant features from the data set, as they can lead to a reduction of the classification accuracy and to an unnecessary increase of computational cost (Blum & Langley, 1997; Koller & Sahami, 1996). Two popular feature filtering methods are InfoGain (IG) (Yang & Pedersen, 1997) and Correlation-based Feature Subset Selection (CfsSubsetEval or CFS) (Hall, 1998). IG measures the amount of information in bits about the class prediction if the only information available is the presence of a single feature and the corresponding class distribution. Practically, IG measures the expected reduction in entropy (uncertainty associated with a random feature) (Mitchell, 1997). CFS measures the “goodness” of a feature subset, considering the usefulness of individual features for predicting the class label along with the level of correlation among them. CFS is based on the hypothesis that good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other (Hall & Smith, 1998).

Recently, there have been several TC innovations. One is the study by Rangel and Rosso (2016), which investigated the impact of emotions on the identification of age and gender. One of their findings is that “females seem to make higher use of emotional verbs (e.g. feel, want, love…) than males, who instead use more language verbs (e.g. tell, say, speak…)”. Investigation of the impact of emotions, as well as other types of sentiment (e.g., Hu, Tang, Gao and Liu (2013); Pang and Lee (2008); Taboada, Brooke, Tofiloski, Voll and Stede (2011)), might help us in future research on complaint classification.

A second innovation is a simple classifier for TC built by Joulin, Grave, Bojanowski and Mikolov (2016). This classifier uses a bag of n-grams, a hashing trick (Weinberger, Dasgupta, Attenberg, Langford & Smola, 2009) and a hashing function applied in Mikolov, Deoras, Povey, Burget and Černocký (2011). In several tasks, this TC classifier is equivalent in quality with deep learning (DL) classifiers in terms of accuracy and is several orders of magnitude faster for training and evaluation. Joulin et al. (2016) showed that although DL methods are considered to have higher representational power than shallow models, shallow models are the right ones to use to evaluate sentiment analysis. The hashing trick and a suitable hashing function might also help us in future studies of complaint classification.

In recent years, more and more TC studies incorporate DL methods within their TC models. For instance, Lee and Dernoncourt (2016) presented sequential short TC (e.g., sentences in a document or utterances in a dialog), using recurrent neural networks (RNNs) (Levin, 1990) and convolutional neural networks (CNNs) (Nowlan and Platt, 1995) that incorporate the preceding short texts. Conneau, Schwenk, Barrault and Lecun (2016) applied very deep char-convolutional networks for TC. They were able to show that increasing the depth up to 29 convolutional layers steadily improves performance over the state-of-the-art results on several public TC tasks. Johnson and Zhang (2016) tested CNNs for TC and compared shallow word-level vs. deep character-level. Their main conclusion for the Yelp.f dataset is that their best shallow word-CNN with 2 layers and 184 M parameters ran for 72 s achieving an error rate of 32.39%, which is much faster than the run time (700 s) of their best char-CNN with 29 + 2 layers and 4.6 M parameters achieving an error rate of 35.28%. . Jiang et al. (2018) proposed a hybrid TC model based on a deep belief network and softmax regression. The experimental results on the Reuters-21,578 and 20-Newsgroup corpora showed that the proposed model performs significantly better than classical algorithms, such as SVM and k-nearest neighbor (kNN) (Whitney & Dwyer, 1966).

As mentioned before, Joulin et al. (2016) showed that in various TC tasks, DL methods are not the right ones to use to evaluate relatively simple TC tasks such as sentiment analysis. This argument is also relevant to our TC task (i.e., classifying complaint letters according to company categories), which is relatively simple. Furthermore, DL methods require a large amount of training data for fine tuning the high number of their parameters (Camilleri & Prescott, 2017; Sun, Tseng, Zhang & Qian, 2017). In our task, the size of

the dataset is relatively small and therefore, it is probably not suitable for DL methods.

2.2. Application of tc to complaint letters

This section summarizes the recent studies in the domain of automatic complaint classification. Jin et al. (2013) distinguished complaints from non-complaints in social media, especially when the number of labeled samples is very small. Given a small set of labeled complaint examples, a small set of labeled non-complaint examples, and a large set of unlabeled examples, their algorithm is applied to enlarge the sets of complaint examples and non-complaint examples. After enough labeled examples have been prepared, SVM and KNN methods are adopted to construct a classifier. Experimental results using SVM indicated the enlargement algorithm significantly increases the performance of TC, especially when the initial number of labeled examples is small. When the percentage of the original labeled examples is 1%, the precision, recall, and F1 values for SVM are 87.9%, 32.8%, and 47.8%, respectively. After the enlargement algorithm is applied, the precision drops to 76.1%, the recall rises to 69.9%, and F1 rises to 72.9%, showing a significant increase in performance.

Choe, Lehto, Shin and Choi (2013) examined the feasibility of extracting useful information from customer comments using an NB classifier. They used a database containing 533 customer calls to call centers in 2009, which was obtained from a Korean mobile telephone service provider, having discarded 150 calls that did not contain any customer complaints or comments. The remaining 383 calls were classified by human experts into four domains and 27 complaint categories. The comments were randomly split into a training set (257 cases, 67%) and a test set (126 cases, 33%). An NB classifier, which was developed on the training set, yielded 75% accuracy for the domain prediction and 51% for the specific subcategory prediction. An analysis of their prediction capability suggests that comments that are difficult to understand by human beings ("complex comments") should be transferred for expert review, while all other comments that can be automatically classified by the classifier are considered "easy comments".

Haifeng and Junhua (2013) divided 1000 customer complaints in the telecommunication industry written in Chinese into seven classifications: broadband, signal, address, internet, order, fee, and other samples. They applied SVM using the LibSVM library (Chang & Lin, 2011) using Term Frequency - Inverse Document Frequency (tf-idf) (Rajaraman & Ullman, 2011) values computed for 897 selected keywords and obtained an accuracy result of 81.4%.

Maia et al. (2014) investigated the application of text mining techniques for automatic classification of complaints for Brazil's CGU (see Section 2). This study presents a model that automatically classifies complaints by using text mining techniques. This model performs various preprocessing techniques: word stemming, stopword removal, low-frequency word removal, conversion of upper to lower case letters, and removal of punctuation, accentuation, numbers, and white spaces. They applied four ML algorithms: SVM, NB, Random Forest (RF) (Breiman, 2001), and Decision Tree (Swain & Hauska, 1977), which were evaluated with various measures including Kappa and F1 (F-measure, F-Score) (Sasaki, 2007). The best results (F1 of 0.84 and a Kappa value of 0.77) were achieved by the RF ML method.

Hayati, Wicaksono and Adriani (2016) proposed several approaches to automatically classify short complaint documents written in Indonesian into nine classes: reformation and governance; education; health; infrastructure; energy and natural resources; environment and disaster management; politics, law, and security; economics; and people's welfare. Their pre-processing algorithm contains removal of stopwords and other noises, and the filtering out of duplicate documents and documents which contain fewer than three words. They applied two ML methods, NB and SVM, using various combinations of word unigrams and word bigrams. For a corpus containing 17,000 short complaint documents, their best accuracy results were 80.89% using 2000 unigrams and 2000 bigrams, and 81.16% using 1500 Latent Dirichlet Allocation (LDA) (Blei et al., 2003) based unigrams, and 2000 bigrams.

Surjandari, Megawati, Dhini and Hardaya (2016) applied text mining methods to classify 27,335 textual complaints written in Indonesian into six classes: equitable access to education; public health; energy, food, and maritime; poverty alleviation; infrastructure development; and bureaucracy reformation. Their pre-processing process contains tokenization, case folding, normalization (substitution of misspelled or abbreviated words), filtering (elimination of special characters), and stemming. They applied the SVM method using tf-idf values. The classification model that used stemming obtained a higher accuracy result (66%) than the accuracy result (53%) of the model without stemming.

Forster and Entrup (2017) presented a cognitive computing approach for classification of dissatisfaction and four complaint-specific complaint classes in correspondence documents between clients and an insurance company. Their approach combines an ME algorithm with language modeling, tf-idf and sentiment analysis. The model was trained and tested with a set of 2500 original insurance communication documents written in German, which were manually annotated by the partnering insurance company. The authors applied a few pre-processing steps. First, the texts were cleaned: headers and footers of emails and letters were removed, and the texts were anonymized. Furthermore, the texts were tokenized, lemmatized, and part-of-speech (POS) tagging was applied. Then the actual features were extracted. In addition to a bag-of-words approach using all lemmas, synonyms were inserted, bigrams extracted, and the lemmas with the highest tf-idf and log-likelihood values were calculated and sorted by their POS tags. Further, each text was assigned a set of topic models, and a sentiment model was used to calculate first value for each word. These values were then combined for overall value for each text document. Some other features were based on pattern matching, e.g., the use of caps-lock and the number of exclamation marks. The binary classification to complaint or no-complaint obtained a precision of 0.84 and a recall of 0.97, resulting in an F1 of 0.90.

3. Methodology

The general methodology of this research is as follows. First, we constructed a corpus of 2073 complaint letters written in Hebrew

that were collected manually from a variety of websites (details in Section 4). These complaint letters belong to six categories as follows: insurance, hospitals, vacation deals, TV, car rental, and communication (further divided into cellular and wired). Each category (or sub-category) contains three companies, with one exception (only two companies).

We decided to represent each document by a BOW vector containing n values representing n selected word unigrams. Each word unigram was represented by its frequency (TF) in the document, normalized by the number of words in the document. This decision was based on our successful experiences with classifying documents written in Hebrew (e.g., HaCohen-Kerner et al., 2016; HaCohen-Kerner, Mughaz, Beck & Yehudai, 2008) as well as on the successful experiences of other researchers with classifying documents written in other languages (e.g., Shahana & Omman, 2015; Yamasaki et al., 2015).

We then constructed a TC model using various versions of word unigrams, four supervised ML methods on the seven categories (insurance, hospitals, vacation deals, TV, car rental, cellular communication, and wired communication), using between 500 (the baseline) and 5000 word unigrams (in increments of 500) to measure the accuracy results.

We considered four popular supervised ML algorithms: Bayes Networks (BayesNet, BN) (Pourret et al., 2008), SimpleLogistic (SL) (Landwehr, Hall & Frank, 2005; Sumner, Frank & Hall, 2005), SMO (Keerthi, Shevade, Bhattacharyya & Murthy, 2001; Platt, 1998), and Random Forest (RF) (Breiman, 2001). These ML algorithms were chosen because they have been found successful for various previous classification tasks (e.g., HaCohen-Kerner, Stern, Korkus & Fredj, 2007; HaCohen-Kerner, Kass, & Peretz, 2010; Gokulakrishnan, Priyanthan, Ragavan, Prasath & Perera, 2012.; Kwon & Sim, 2013). A brief description of each one of these algorithms is given in the next paragraph.

BN is a variant of a probabilistic statistical classification model, which represents a set of random variables and their conditional dependencies via a directed acyclic graph. SL is a variant of Logistic regression (LR) (Cessie & Van Houwelingen, 1992). SL, which is implemented in the WEKA platform, implements a multinomial LR model with a ridge estimator. SMO is a variant of the SVM method. SMO is an iterative method created to solve the optimization problem frequently found in SVM methods. SMO divides this problem into a series of the smallest possible sub-problems, which are then resolved analytically. RF is an ensemble learning method for classification and regression. RF operates by constructing a multitude of decision trees at training time and outputting classification for the case at hand. RF combines Breiman's "bagging" (Bootstrap aggregating) idea (Breiman, 1996) and a random selection of features introduced by Ho (1995) to construct a forest of decision trees.

The combination with the best accuracy result (best supervised ML method, best set of word unigrams, and the best version with/without stopwords) was re-applied using two feature filtering methods: IG and CFS (see Section 2.1). Then, we performed parameter tuning to the best combination of the ML method, with/without stopwords (see next paragraph), several features, and best feature filtering method.

We also introduce a semantic analysis of the experimental results using various defined semantic fields (as mentioned before, a 'semantic field' is a group of words with a common core of meaning), facilitating a deeper analysis of the characteristics of the complaint letter in each domain. The semantic analysis was performed on the features that led to the best accuracy result for each of the final classification experiments from seven to four categories. Due to the semantic analysis, important, interesting, and sometimes surprising findings were found for all the categories in general, and for one specific category in particular. These findings can help service providers to improve their service and to focus on the important issues.

4. A corpus of complaint letters

In order to create a corpus of complaint letters we contacted 26 companies in various categories. Then based on the results we determined the following six company categories*: communication, TV, insurance, vacation deals, car rental, and hospitals*, where the *communication* category was further divided into two sub-categories: *wired communication and cellular communication*. Each category (or sub-category) contains three companies, except for one category that includes only two companies.

Initially, we approached these 26 companies and asked them for complaint letters that had been sent to them. Most of the companies (18) did not reply. Seven companies declined our request on the grounds that access to the letters would be harmful to customer confidentiality. Only one company took the trouble to sent us complaint letters and even this company sent us only less than ten letters. As a result, we decided to search for letters of complaint on websites in the Internet.

The letters were collected manually from a variety of sites. The three main sites were:

(1) www.tluna.co.il – This website features letters of complaint addressed to a wide range of companies. The site advocates transferring the letter to the company and demands the company's response and treatment. A total of 1347 letters were collected from this source.
(2) Facebook – www.facebook.co.il. There are many groups and pages from people who have been harmed by various companies. There, they raise their issues regarding a given company. A total of 1007 letters were collected from this source.
(3) www.sherut.net – is another site that collects letters of complaint about various companies. A total of 303 letters were collected from this source.

Note that many of the letters are not written correctly in terms of the proper structure of a general letter (the letters are not clearly divided into three components: introduction, body, and conclusion). For example, some of the letters include only one paragraph and some of them only include one sentence (which is often longwinded and cumbersome).

We collected 2434 complaint letters, from which we removed 361, mainly because they referred to companies that each had a relatively low number of complaint letters (e.g., one hospital with only two complaint letters and two insurance companies with only

Table 1
The constructed corpus.

Category	Name of company	# of complaint letters	Total # of complaint letters per category
Wired communication	Net vision	205	475
	012smile	167	
	Bezeq	103	
Cellular communication	Partner	146	423
	Cellcom	135	
	Pelephone	142	
TV	Yes	210	410
	HOT	200	
Insurance	9,000,000	108	295
	AIG	95	
	555	92	
Vacation deals	Issta	88	209
	Daka90	71	
	Flying Carpet	50	
Car rental	Shlomo Sixt	90	171
	Avis	45	
	Budget	36	
Hospitals	Ichilov	35	90
	Kaplan Medical Center	33	
	Assuta	22	
Total # of complaint letters			2073

27 and 13, respectively). In our study, we wanted to classify according to categories of companies (e.g., cellular communication, insurance companies, and vacation deals). Therefore, we wanted each category to contain at least 2–3 different companies so that the classification would be according to general categories rather than specific companies. For this reason, we removed 48 complaint letters of one company, which had a monopoly in its domain of activity in Israel (the Israel Electric Company).At the end of the process, there were 2073 complaint letters from 20 companies that were divided into six categories as follows: insurance, hospitals, vacation deals, TV, communication, and car rental. These six categories can be considered seven categories, since the communication category can be divided into two sub-categories: cellular communication and wired communication.

Table 1 presents the corpus we constructed. This table contains the names of the seven categories (the two communication sub-categories are considered two categories), the names of the companies, and the number of complaint letters that belong to each company. In addition, we present the total number of complaint letters per category and the total number of all complaint letters.

In Appendix A, we present several examples taken from letters of complaint from the insurance, communication, and vacation deals categories.

5. Model and experimental results

Below, we present a flowchart of our classification model that provides classification of complaint letters from $N = 7$ to $N = 3$ using the following settings: Train (67%)/Test (33%) (data randomized) and the number of repetitions of the experiment is 10 (Fig. 1).

As mentioned before, the initial runs were performed with and without two versions of stopwords. We prepared two lists of stopwords for our experiments. The first list of stopwords – *SW1* – the basic list of stopwords contains 48 general Hebrew words (e.g., and, after, but, if, or, they). These 48 stopwords were among the top occurring word unigrams in the complaint letters in all the seven categories. These stopwords are presented in the basic list of stopwords (SW1, see Table 15 in Appendix B). About 30 words were taken from HaCohen-Kerner and Blitz (2010), and the additional general words were carefully added by the authors of this paper after manually checking the top occurring word unigrams in the training dataset.

We also observed that 190 words were company names, category names, their nicknames, their initials, and the same words with different prefix combinations in Hebrew. According to linguistic estimates, the Hebrew language is richer than English in its morphology forms (Wintner, 2004). Hebrew has about 70,000,000 valid (inflected) forms, while English has only about 1000,000. In Hebrew, there are up to 7000 declensions for only one stem, while in English, there are only a few declensions. A word in Hebrew might include one or more prefix letter(s and/or suffix letter(s), where each letter represents a single word in English. For example, the single Hebrew word "וכשיאכלוהו" (vkhshy'khluhu)[2] is translated into the following sequence of six English words: "and when they will eat it" where the first Hebrew letter "ו" is translated into "and" and the two next Hebrew letters "כש" are translated "when". HaCohen-Kerner and Erlich (2014) presented a method to find whether a certain word in Hebrew includes in its beginning a known prefix.

Then, we added these 190 words to SW1, resulting in SW2 – the extended list of stopwords – containing 238 words. We also

[2] The Hebrew Transliteration Table, which has been used in this paper, is taken from the web-site of the Princeton university library (http://library.princeton.edu/departments/tsd/katmandu/hebrew/trheb.html). Last access: 19/APR/19/

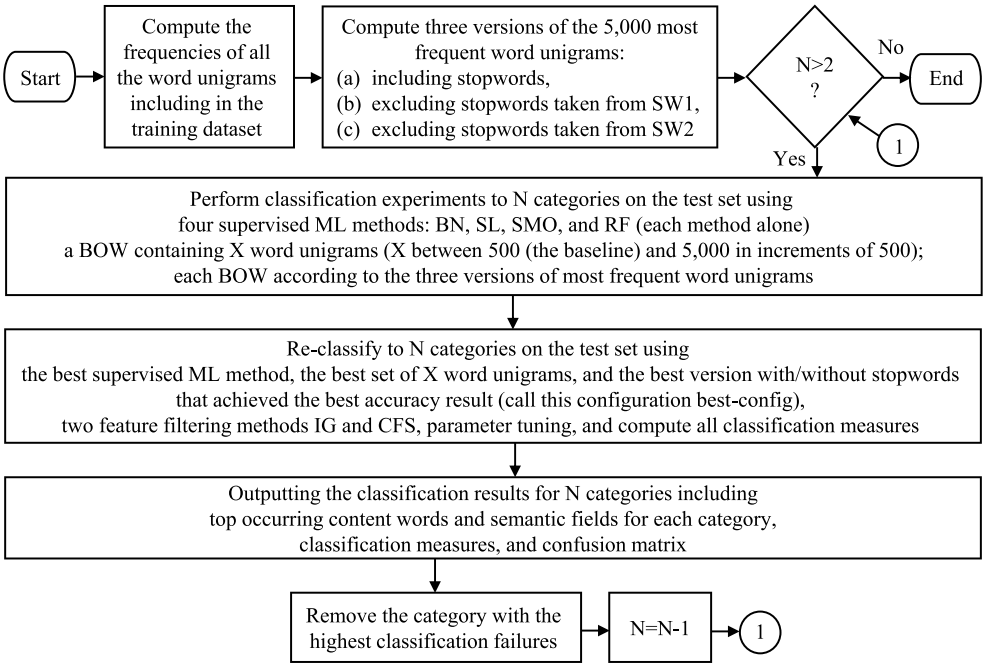

Fig. 1. Flowchart of our classification model.

performed classification experiments excluding SW2, since it might be "too easy" to classify the complaint letters to the seven categories when such unique and distinct data are included in the complaint letters.

We used the WEKA platform with its default parameters (Hall et al., 2009; Witten et al., 2005). For each TC task, we used the experimenter mode in WEKA Version 3.9.1 with the following settings: Train (67%)/Test (33%) (data randomized) and the number of repetitions of the experiment was set to 10.

The results in the following tables (classification experiments to seven to four categories) are according to the accuracy measure. For the best result in each experiment, we provided results of additional measures derived from parameter tuning: precision, recall, F1, a precision-recall curve (PRC-area), a plot of precision vs. recall (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015), and a receiver operating characteristics curve (ROC-area) (a plot of true positive fraction vs. false positive fraction) (Hanley &, McNeil,1982; Bradley, 1997; Provost, Fawcett & Kohavi, 1998; Fawcett, 2006) and more insights and analyses.

The tables in this article contain various annotations and colors. The annotation v or * indicates that a specific result in a certain column is statistically better (v) or worse (*) than the baseline result (i.e., the result with 500 word unigrams) at the significance level of 0.05. The blue color represents accuracy results that were significantly better than the baseline result. The purple color represents the highest accuracy result obtained by a certain ML method (each ML method is represented by another column). The red color represents the best accuracy result in the whole table.

5.1. Classification experiments to seven categories

We performed three general classification experiments in the seven categories (insurance, hospitals, vacation deals, TV, cellular communication, wired communication, and car rental): one including stopwords, the second by eliminating the stopwords in SW1, and the third without the stopwords included in SW2. Tables 2-4 present the accuracy results of these classification experiments using various sets of the most frequent word unigrams, including stopwords, excluding stopwords from the basic list of stopwords (SW1), and excluding stopwords from the extended list of stopwords (SW2), respectively.

The main findings derived from Table 2 are: (1) the best accuracy result (81.56%) was obtained by SL using 5000 word unigrams and (2) the SL method was the best ML method, and all its results significantly improved the baseline result (75.03%) using 500 word unigrams.

The main findings derived from Table 3 (while excluding SW1's stopwords) are similar to those of Table 2, but slightly better. The results are as follows: (1) the best accuracy result (81.59%) was obtained by SL using 4500 word unigrams and (2) the SL method was again the best ML method, and all its results significantly improved SL's baseline result (75.82%) using 500 word unigrams.

The main findings derived from Table 4 (while excluding SW2's stopwords) are: (1) the accuracy results are significantly lower than those achieved in Tables 2 & 3 and (2) the best accuracy result (76.04%) was obtained by BN using 3000 word unigrams.

Table 5 presents the best accuracy results of the classification experiments to seven categories for each ML method, including stopwords, excluding SW1's stopwords, and excluding SW2's stopwords.

The main findings derived from Table 5 are: (1) The best accuracy result (81.59%) was obtained by SL using 4500 word unigrams

Table 2
Classification experiments to seven categories using word unigrams (including stopwords).

Features	SMO	RF	SL	BN
500	69.55	75.02	75.03	70.64
1,000	73.4 v	75.46	78.45 v	74.92 v
1,500	75.7 v	**75.76**	79.53 v	76.32 v
2,000	77.3 v	74.26	80.72 v	77.49 v
2,500	77.11 v	73.02	80.03 v	77.79 v
3,000	**77.49 v**	72.55*	80.28 v	77.81 v
3,500	77.39 v	71.29*	80.99 v	77.88 v
4,000	77.37 v	71.32*	81.49 v	77.88 v
4,500	76.83 v	70.34*	81.34 v	**77.9 v**
5,000	76.49 v	70.07*	**81.56 v**	77.88 v

Table 3
Classification experiments to seven categories using word unigrams (without SW1).

Features	SMO	RF	SL	BN
500	70.46	**77.93**	75.82	75.41
1,000	**73.48 v**	**78.61**	**78.53 v**	**78.76 v**
1,500	**76.6 v**	77.9	**79.72 v**	**79.99 v**
2,000	**77.3 v**	77.17	**80.77 v**	**80.7 v**
2,500	**77.3 v**	77.05	**81.11 v**	**81.08 v**
3,000	**77.37 v**	76.19	**81.29 v**	**81.14 v**
3,500	**77.14 v**	74.46	**81.15 v**	**81.23 v**
4,000	**77.27 v**	73.47*	**81.17 v**	**81.18 v**
4,500	**76.93 v**	73.37*	**81.59 v**	**81.2 v**
5,000	**77.03 v**	72.99*	**81.56 v**	**81.21 v**

Table 4
Classification experiments to seven categories using word unigrams (without SW2).

Features	SMO	RF	SL	BN
500	64.63	70.48	68.21	68.98
1,000	67.85 v	70.45	70.91	72.71 v
1,500	70.59 v	**71.23**	72.58 v	74.2 v
2,000	71.44 v	69.72	72.59 v	75.53 v
2,500	71.91 v	69.25	72.53	75.85 v
3,000	72.04 v	68.58	72.43 v	**76.04 v**
3,500	72.04 v	66.4*	73.27 v	76.01 v
4,000	72.15 v	67.03*	.73 06 v	75.97 v
4,500	**72.39**	66.37*	73.27 v	75.97 v
5,000	71.73 v	65.77*	**73.75 v**	75.94 v

Table 5
Best accuracy results for each ML method with/without various stopword list versions.

ML Method	including stopwords	without SW1	without SW2
Best SMO	**77.49**	77.37	72.39
Best RF	75.76	**78.61**	71.23
Best SL	81.56	**81.59**	73.75
Best BN	77.9	**81.21**	76.04
Best result	81.56	**81.59**	76.04

excluding SW1. This result was slightly better than the second best result (81.56%) obtained also by SL using 5000 word unigrams including stopwords. (2) In most of the ML methods, the best result was achieved excluding SW1. (3) In all the ML methods, the results obtained excluding SW2 were significantly worse than the results obtained with the two other versions of stopwords. This finding can be explained by the fact that SW2 contains names of companies, their nicknames, their initials, including commonly used prefixes. These words certainly help to classify according to categories. The removal of such words significantly decreases the accuracy results.

As mentioned, the best result from the three experiments was achieved excluding SW1 by three out of four ML methods. Because we planned an extensive set of experiments, we decided to perform all the remaining classification experiments excluding SW1. By removing the stopwords in SW1 from the dataset, we could get more words relevant to the letters and the fields themselves.

This decision is consistent with the fact that many TC systems perform classification experiments excluding stopwords. For instance, Forman (2003), in his study on feature selection metrics for TC, claimed that stopwords are ambiguous and occur frequently enough that they do not discriminate for any particular class. Stopword removal improves classification accuracy results in many TC systems (e.g., Song et al., 2005; Gonçalves, Gonçalves, Camacho & Oliveira, 2010; Méndez, Iglesias, Fdez-Riverola, Díaz & Corchado, 2005).

5.2. Feature filtering and parameter tuning

We performed a series of experiments to improve the best result using feature filtering and parameter tuning. First, we performed various filtering experiments in order to improve the baseline result, i.e., the best accuracy result 81.59% in Table 5 that was achieved by SL using 4500 word unigrams excluding SW1. The best accuracy result (84.51%) was achieved with the top 700 word unigrams according to IG. This was a significant improvement of about 3% to the baseline result. Additional experimental results were as follows: All the filtering versions of IG improved the baseline result. However, only the versions that used at least 500 top word unigrams significantly improved the baseline result. The result obtained using the CFS filtering method was significantly worse than the baseline result.

Parameter tuning was performed to further improve the best result (84.51%), achieved with feature filtering. The values of the modifiable parameters heuristicStop (default value: 50), maxBoostingIterations (default value: 500), numBoostingIterations (default value: 0), and weightTrimBeta (default value: 0.0) were altered with 14 different approaches, none of which improved the results. We tried 14 trial and error attempts to change the values of one or more of these four parameters (using both lower and higher values than the default values), e.g. heuristicStop (values of 5, 500, and 5000), maxBoostingIterations (values of 50 and 5000), numBoostingIterations (values of 1 and −1), and weightTrimBeta (values of −0.5, −1, 0.5, and 1). All the attempts failed to improve the result mentioned above.

Semantic analysis of the features that led to the best accuracy result for seven categories is given in Section 6.1.

Table 6 presents the confusion matrix for the classification experiment to seven categories that was performed by SL with its best setting.

Analysis of the results presented in Table 6 shows that the classification quality for the communication categories (both landline and cellular) is the lowest, and most of the incorrect classifications happened in these two types of communication.

This phenomenon can be explained by the fact that both categories relate to communication and there are no sufficient unique characteristics for each category alone. Nonetheless, the great majority of communication letters were classified as communication categories and not as other categories, which teaches us that the categories of communication have many unique characteristics.

In addition, we see that a considerable number of cases (36) in the TV category were classified as communication. Television is indeed a type of communication, which explains these incorrect classifications. We also see that the classification quality for hospitals is not high. This finding might be because there are not enough distinct characteristics for hospitals, especially since this category is the smallest in terms of its number of complaint letters (only 90).

The cellular communication category was the category with the most classification failures (both in terms of number (105) and its relative percentage of 25.4%). Therefore, we decided to filter out the instances of this category and perform classification experiments with the other six categories, which we found to be more successful in terms of their classification quality. As shown above, in each stage we can see which category has the lowest accuracy rate, and derive possible explanations for the difficulty of classifying this category, as opposed to the other categories. Running experiments on fewer categories enables us to identify which categories are more difficult to classify and provide reasonable explanations for these findings.

Table 6
Confusion matrix for 7 categories.

a	b	c	d	e	f	g	← classified as
234	1	1	7	8	15	29	*a* = insurance
3	69	0	1	2	10	5	*b* = hospitals
3	1	178	3	2	10	12	*c* = vacation deals
6	1	5	148	0	5	6	*d* = car rental
2	0	4	0	368	19	17	*e* = TV
9	3	3	3	14	318	73	*f* = cellular communication
14	3	3	2	19	49	385	*g* = wired communication

Table 7
Classification experiments to six categories with word unigrams (without SW1).

Features	SMO	RF	SL	BN
500	77.91	**83.24**	82.76	81.09
1,000	80.99 v	83.02	84.85 v	83.72 v
1,500	82.1 v	83.2	85.9 v	84.78 v
2,000	82.48 v	81.77	86.43 v	85.02 v
2,500	82.83 v	81.14	85.99 v	**85.24 v**
3,000	**83.44 v**	79.22*	**86.71 v**	85.2 v
3,500	83.05 v	78.7*	86.16 v	85.13 v
4,000	83.42 v	77.62*	86.38 v	85.07 v
4,500	83.36 v	77.97*	85.81 v	85.07 v
5,000	82.78 v	76.55*	86.56 v	85.13 v

5.3. Classification experiments to six categories

In this Section, we briefly describe the classification experiments to six categories after removing the instances of the cellular communication category. Table 7 presents the results of these experiments with various numbers of word unigrams and without SW1.

For six categories as well, we conducted a series of experiments to improve the best result using feature filtering and parameter tuning. The best accuracy result (88.38%) was achieved with the top 659 word unigrams according to IG, which is an improvement of 1.67% compared to the best result in Table 7 (86.71%). Using parameter tuning, we further improved to 88.41%, yielding a total improvement of 1.7% from 86.71%. Semantic analysis of the features that led to the best accuracy result for 6 categories is given in Section 6.2.

Table 8 presents the confusion matrix for the classification experiment to 6 categories that was performed by SL with its best setting.

Analysis of the results introduced in Table 8 shows that the classification quality for the insurance category is the lowest. A possible explanation might be that there are not enough distinct characteristics for insurance.

5.4. Classification experiments to 5 categories

In this Section, we briefly describe the classification experiments to 5 categories after removing the instances of the insurance category. Table 9 presents the results of these experiments with various numbers of word unigrams and without SW1.

For five categories as well, we conducted a series of experiments to improve the best result using feature filtering and parameter tuning. Applying SL using the top 400 word unigrams according to IG yields an accuracy result of 91.19%, an improvement of 1.72% compared to the best accuracy result 89.47% in Table 9. Using parameter tuning that changes heuristicStop from 50 to 500, we achieved a small improvement of 0.02% to 91.21% presenting a total improvement of 1.74% from 89.47%. Semantic analysis of the features that led to the best accuracy result for 5 categories is given in Section 6.3.

5.5. Summary of classification experiments between seven and four categories

We next reduced the number of categories to four. In Table 10, we compare the best results (using SL, the best ML method in our experiments, feature filtering using IG, and parameter tuning) for different numbers of categories (seven, six, five, and four).

Table 10 shows that the fewer the categories, the higher the accuracy rate. The best improvements appeared when we reduced the number of different categories from seven to six (3.9%) and from six to five (2.8%). In general, the improvements occurred along all the examined measures: accuracy, precision, recall, F1, PRC area, and ROC area.

Table 8
Confusion matrix for 6 categories.

a	b	c	d	e	f	← classified as
237	2	2	5	8	41	*a* = insurance
3	76	0	0	4	7	*b* = hospitals
4	2	190	1	3	9	*c* = vacation deals
7	0	2	153	1	8	*d* = car rental
5	0	1	2	362	40	*e* = TV
23	3	1	1	19	428	*f* = wired communication

Table 9
Classification experiments to 5 categories with word unigrams (without SW1).

Features	SMO	RF	SL	BN
500	82.8	**87.61**	87.32	84.82
1,000	85.08	86.31	88.26	86.54 v
1,500	86.45v	85.4	**89.47 v**	86.94 v
2,000	86.65 v	84.3*	89.09 v	87.1 v
2,500	87.18 v	82.71*	88.8	**87.19 v**
3,000	86.81 v	81.64*	89.09	87.05 v
3,500	**87.76 v**	80.61*	88.91	87.05 v
4,000	86.25 v	79.76*	89.18	87.07 v
4,500	85.91 v	79.85*	89.42	87.07 v
5,000	85.66 v	77.52*	89.22	87.07 v

Table 10
Comparison of the best results in the different categories.

# of categories	Without the following categories	# of complaint letters	The best accuracy result	Precis-ion	Recall	F1	PRC area	ROC area
7	–	2073	**84.51**	0.85	0.8	0.82	0.91	0.97
6	Cellular communication	1650	**88.41**	0.84	0.84	0.84	0.93	0.98
5	Insurance	1355	**91.21**	0.92	0.84	0.88	0.94	0.99
4	Hospitals	1265	**91.86**	0.97	0.89	0.93	0.97	0.99

6. Semantic analysis of selected results

In this section, we introduce a semantic analysis of many selected results from the classification experiments to five to seven categories.

6.1. Analysis of the best accuracy result for seven categories

Table 11 presents the top 34 words according to best accuracy result (84.51%) that was achieved by SL with the top 700 word unigrams according to IG for seven categories. These top words are presented in Hebrew together with their translations into English,

Table 11
Top 34 words according to IG for seven categories.

#	The word in Hebrew	Trans-literation[a]	translation into English	weight	#	The word in Hebrew	Trans-literation	translation into English	weight
1	רכב	ʁexev	car	0.28	18	סלקום	selkom	Cellcom (a company name)	0.1
2	הרכב	ha-ʁexev	the car	0.23	19	טכנאי	texnai	technician	0.1
3	ביטוח	bituax	Insurance	0.19	20	פלאפון	pelefon	Pelephone (a company name)	0.1
4	הוט	hot	Hot (a company name)	0.18	21	למלון	la-malon	to the hotel	0.1
5	????[b]		????	0.17	22	שלכם	ʃelaxem	your	0.1
6	?????		?????	0.17	23	ההזמנה	ha-hazmana	the booking	0.09
7	??????		??????	0.17	24	במלון	ba-malon	at the hotel	0.09
8	???		???	0.17	25	סיקסט	sikst	Sixt (a company name)	0.09
9	??		??	0.16	26	החולים	ha-xolim	patients	0.09
10	יס	yes	Yes (a company name)	0.15	27	????????		????????	0.09
11	???????		???????	0.14	28	לי	li	to me	0.08
12	הטיסה	ha-tisa	the flight	0.13	29	אני	ani	I	0.08
13	הביטוח	ha-bituax	insurance the	0.12	30	נופש	nofeʃ	vacation	0.08
14	טיסה	tisa	flight	0.12	31	בזק	bezek	Bezeq (a company name)	0.08
15	המלון	ha-malon	the hotel	0.12	32	הממיר	Ha-memiʁ	the converter	0.08
16	איסתא	ista	Issta (a company name)	0.11	33	להתנתק	lehitnatek	to be disconnected	0.08
17	מלון	malon	hotel	0.11	34	חולים	xolim	patients	0.08

[a] The words were transliterated phonetically according to IPA (International Phonetic Alphabet) rules.
[b] Various top occurring tokens are sequences of different lengths of question marks.

Table 12
The distribution of the series of question marks normalized by the number of letters in each domain across the seven categories.

Category # of appearances	Insurance	hospitals	vacation deals	car rental	TV	cellular communica-tions	wired communica-tions
??	0.31	0.344	0.287	0.116	0.19	0.366	0.31
???	0.128	0.2	0.181	0.058	0.114	0.215	0.128
????	0.081	0.122	0.071	0.017	0.058	0.073	0.081
?????	0.03	0.088	0.047	0.006	0.029	0.033	0.03
??????	0.016	0.066	0.038	0	0.021	0.021	0.016

their transliterations, and their weights according to IG.

The main findings and conclusions derived from Table 11 are: (1) most of the words are relevant to specific fields, such as car rental, insurance companies, and vacation deals; (2) additional words are names of specific companies (e.g., Cellcom, Hot, Issta, Bezeq, and Sixt). These words are valuable in identifying the field to which they belong; and (3) additional top occurring tokens are pronouns (e.g., "I" and "me") and sequences of different lengths of question marks (e.g., "??","???","????", and "?????") that surprisingly were helpful.

The important and surprising contribution of question mark series to the success of the classification led us to investigate the distribution of these series across the seven categories. Table 12 introduces the distribution of these series across the seven categories by values of their appearances normalized by the number of letters in each domain across the seven categories.

The main findings and conclusions derived from Table 12 are: (1) in general, in most categories, series of question marks appear in lengths of two to six at a relatively high frequency; (2) the question mark series are distributed unevenly among the categories. The general picture is that categories such as hospitals, cellular communication, and wired communication use these series often, and there are categories, such as car rental and television, which rarely use these series. (3) Relatively speaking, the two-question-mark sequence is most prominent in the communication categories (both wired and cellular). This type of sequence appears the least in car rental. (4) The three-question-mark sequence appears the most in the domain of cellular communication, followed by hospitals. This feature appears the least in the domain of car rental. (5) The four-question-mark sequence appears most in the domain of hospitals, followed by insurance. (6) The five- and six- question-mark sequences are most prominent in hospitals, followed by vacation deals.

A possible explanation for the high number of question mark series in the categories of hospitals and communication: we assume that a series of several question marks is an emotional substitute for the letter author's demand for an immediate answer. The greater the number of question marks in the series, the more urgency to receive a response is expressed. Perhaps this phenomenon expresses more anger, and a greater demand for an immediate response.

Another possible explanation is that the high number of question mark series indicate greater frustration of the writers. Hospitals are the most emotional domain, and there is a greater expectation of receiving quality treatment. It is possible that in more commercial categories (e.g., communication, car rental, and hotels), customers better understand that the service is more technical and less attentive to their welfare, and therefore their expectations are lower. These findings can attest to the intensity in the complainant's attitude towards the object of the complaint: From the hospitals, he expects more, so his frustration that his expectations are being ignored is greater. In the communication categories, the expectation is also great (every fault sabotages his daily functioning). Therefore, in the communication categories, there is also a great deal of frustration (expressed by many question marks). In other domains, frustration is expressed and urgency is displayed, but at reduced levels. A further and deeper analysis of the questions mark series and their interpretation can be done in future research using sentiment analysis tools and methods.

The most frequent content words for each category were found automatically by our program. They are listed below according to the seven categories.

Insurance: insurance, already, vehicle, service, policy, company, directly, days, insured party, number, apartment, get, simple, do, day, again, customers, years, policy, cancel, price, hour, customer, price, representative, month, need, case, good, contact, call, month, duty, want, confirmation, insurer, ask, minute, sum.

Hospitals: patients, surgery, home, receive, there, pay, day, note, say, again, do, time, emergency room, nurse, room, reach, doctor, hour, nose, test, need, treatment, date, number, contact, can, department, home, boy, heart, answer, receive, want, pass, blood, child.

Vacation deals: hotel, flight, road, day, hour, order, date, say, company, service, there, note, contact, talk, trip, morning, number, contact, receive, thing, be said, do, site, request, aviation, representative, order, deal, vacation, return, vacation, room, place, simple, reach, customer, minute, hello, cancel, price, stars, pay.

Car rental: vehicle, company, day, number, receive, date, say, branch, pay, road, there, customer, contact, garage, insurance, talk, service, call, do, address, time, note, hour, hello, credit, need, bring, month, arrive, manager, damage, directly, private, answer, telephone.

Television: technician, representative, customer, day, contact, service, arrive, company, request, minute, number, home, disconnect, month, date, telephone, say, converter, receive, there, hello, call, representative, equipment, talk, hour, hot, month, problem, pay, install, wait, address, year, note.

Cellular communication: device, service, cell phone, receive, day, customer, number, company, representative, telephone, month, request, call, there, say, want, call, pay, do, new, contact, minute, note, representative, year, manager, sim, reach, charge, time, repair, line.

Table 13
Most frequent content words for each category.

#	Insurance	Hospitals	Vacation deals	Car Rental	Television	Cellular Communica-tions	Wired Communica-tions
1	insurance	patients	hotel	vehicle	technician	device	service
2	already	operation	flight	company	representative	service	internet
3	vehicle	home	road	day	customer	cell phone	technician
4	service	receive	day	receive	day	receive	day
5	policy	doctor	hour	date	contact	day	representative
6	company	pay	order	say	service	customer	customer
7	direct	day	date	branch	arrive	number	second
8	days	note	say	pay	company	company	receive
9	insured party	say	company	road	request	representative	home
10	apartment	again	service	there	second	telephone	number
11	receive	do	there	customer	disconnect	month	contact
12	do	time	note	contact	month	request	company
13	day	emergency room	contact	garage	date	contact	contact
14	again	nurse	talk	insurance	telephone	say	do
15	customers	room	trip	talk	say	want	line
16	years	arrive	morning	service	converter	call	mega
17	cancel	doctor	contact	contact	receive	pay	hour
18	price	hour	receive	do	hello	month	telephone
19	hour	nose	be said	address	contact	address	pay
20	customer	test	do	time	call	second	price
21	price	need	site	note	representative	contact	month
22	representative	treatment	request	hour	equipment	note	talk
23	month	date	aviation	credit	talk	year	thank you
24	contact	number	representative	must	hour	manager	arrive
25	contact	address	order	bring	hot	sim	representative
26	obligation	able	deal	month	month	arrive	request
27	want	department	vacation	arrive	problem	charge	disconnect
28	confirmation	to the home	return	manager	pay	repair	month
29	insurer	boy	vacation	damage	installation	line	infrastructure
30	request	heart	room	directly	waiting		year

Wired communication: service, internet, technician, day, simple, representative, customer, minute, receive, home, number, connection, company, call, talk, do, line, mega, hour, telephone, pay, price, month, talk, thank you, reach, request, disconnect, month, infrastructure, year.

These content words were manually analyzed by our linguistic expert and divided into various semantic fields. A 'semantic field' is a group of words with a common core of meaning. A semantic field contains all the words that relate to their common concept, regardless of whether or not they share the same root or other morphological-external identification mark (Brinton, 2000). Anthropologists have employed the method of classification by semantic fields to describe how the tribes and cultures they study envision the world, based on an assumption that this linguistic analysis can testify as to cultural differences (Sovran, 1994).

According to Seker (1989), tracing ways of using words that comprise a field facilitates an effective presentation of a picture of a world (both physical and conceptual) and provides a range of ways to explain its features. The relationships between words within the field (internal hierarchy, the guiding principles for placing them in this particular field, etc.) make it possible to cross-check information from various directions to focus meanings and delineate attributes. These attributes are evidence of a collective psychological or socio-cultural repertoire (Pennebaker, Matthias, Mehl & Niederhoffer, 2003).

Based on Gergen and Gergen (1987), the most frequent words were manually divided by our linguistic expert into semantic fields: service provider-customer relations, price, time, location, and subject domain (i.e., words that are related to the domain itself; for instance, for the "hospital" domain words that relate to the patient, the medical treatment, and the medical staff).

Table 13 presents the top occurring 30 content word unigrams for each category. The words (not necessarily single words) listed in Table 13 are in English and they are translations from single words in Hebrew.

We will define the following colors for the five most important semantic fields:

- Yellow – Words related to the subject category
- Purple – Words related to time
- Red – Words related to location

Table 14
Distribution of general semantic fields among the seven categories.

Category	service provider - customer relations	subject domain	time	price	location
Insurance	14	7	7	2	
Hospitals	9	12	5	1	1
Vacation deals	15	8	4		3
Car Rental	16	4	5	2	3
Television	19	4	6	1	
Cellular communication	13	8	5	2	
Wired communication	14	7	6	2	1

- Blue – Words related to the service provider - customer relations
- Grey – Words related to price

Table 14 introduces the distribution of semantic fields among the seven categories. The values listed in the cells are the frequencies of these semantic fields in the different categories (30 most common in each category).

The analysis of the results shown in Table 14 indicates common characteristics of the complaint letters in each domain, as well as the subjects that concern customers in the letters, and whether there is significance to the frequency or infrequency of the use of various semantic fields in each domain.

The main findings and conclusions derived from Table 14 are: in all categories, apart from the 'hospitals' category, the most important (frequent) semantic field is "service provider – customer relations". These findings are compatible with the findings of Pfeil, Yersin, Trueb, Feiner and Carron (2018), who examined complaint letters sent to the Swiss Emergency Department. They found that the complaints mainly concerned medical care, organizational aspects, and staff-patient communication.

Translated examples for complaints that are relevant to the semantic field of "service provider – customer relations", which are relevant to all categories are: (1) "There is no doubt that in advertising you are huge, but in customer service, unfortunately, you need a significant improvement"; and (2) "What does all the rebranding help if your service gets worse?"

The second most frequent semantic field is "subject domain", which is widely used in the categories of 'insurance', 'vacation deals', 'cellular communication' and 'wired communication'. In the category of 'hospitals', "subject domain" is the most important (frequent) semantic field.

Translated examples for complaints that are relevant to the semantic field of "subject domain" are: (1) "The doctor talked to us last Tuesday about doing the surgery … they claimed that there is not enough manpower!!!! Who cares? This is a woman who has been in intensive care for two weeks … Why this criminal neglect?" (category of hospitals); (2) "We were in a 4-star hotel …, believe me, friends, 1 star is a lot" (category of vacation deals); and (3) "Overseas calls to numbers overseas did not work in the countries I was in" (category of cellular communication).

The third most important semantic field among all the categories is the "time" field. Less important semantic fields are "price" and "location". Translated examples for complaints relevant to these three semantic fields are, respectively: "I have been trying to call for a whole week and wait more than ten minutes, and no representative answers me and does not get back to me" (category of cellular communication); (2) "Comprehensive insurance, Are you sitting? 12,000 NIS? And now in words: Twelve thousand Shekels!!!" (category of insurance); and (3) "Unfortunately, the driver took us only to about 750 m away from the hotel … and we had to walk

Table 15
SW1 – Basic list of stopwords (contains 48 words) in Hebrew.

#	word in Hebrew	trans-literation	translation into English	#	word in Hebrew	trans-literation	translation into English	#	word in Hebrew	trans-literation	translation into English
1	אבל	aval	but	17	הם	hem	they	33	כל	kel	all
2	או	o	or	18	ו	ve	and	34	כש	xʃe	when
3	אחר	axaʁ	after	19	וה	Ve-ha	and the	35	ל	l	to
4	אחרי	axaʁei	after	20	וכ	ux	and about	36	לא	lo	no
5	אך	ax	but	21	וכש	uxʃe	and when	37	לאחר	leahar	after
6	אם	im	if	22	ול	Ve-le	and to	38	ליד	leyad	near
7	את	et	term used to indicate a direct object	23	ולכש	Ve-lixʃe	and when	39	לפני	lifney	before
8	ב	be	in	24	זאת	zot	this	40	מעל	meal	above
9	בין	bein	between	25	זה	ze	this	41	מתחת	mitahat	under
10	גם	gam	too	26	זו	zo	this	42	עד	ad	until
11	ה	ha	the	27	יותר	joteʁ	more	43	עוד	od	more
12	הוא	hu	he	28	יחד	jaxad	together	44	על	al	on
13	היא	hi	she	29	יש	jeʃ	there is	45	עם	im	with
14	היה	haja	was	30	כ	ke	as	46	ש	ʃ	that
15	היו	haju	there were	31	כי	ki	because	47	שה	ʃe-ha	that the
16	הייתה	hajta	was	32	כך	kax	so	48	של	ʃel	of

Table 16
Top 34 words according to IG for six categories.

#	word in Hebrew	trans-literation	translation into English	weight	#	word in Hebrew	trans-literation	translation into English	weight
1	רכב	ʁexev	car	0.31	18	טכנאי	texnai	technician	0.13
2	הרכב	ha-ʁexev	the car	0.26	19	שלכם	ʃelaxem	your	0.11
3	ביטוח	bituax	insurance	0.24	20	למלון	la-malon	to the hotel	0.11
4	הוט	hot	Hot (a company name)	0.2	21	במלון	ba-malon	at the hotel	0.1
5	????		????	0.18	22	החולים	ha-xolim	patients the	0.1
6	?????		?????	0.18	23	ההזמנה	ha-hazmana	the booking	0.1
7	??????		??????	0.18	24	סיקסט	sikst	Sixt (a company name)	0.1
8	???		???	0.18	25	אני	ani	I	0.09
9	??		??	0.17	26	לי	li	to me	0.09
10	יס	yes	Yes (a company name)	0.16	27	בזק	bezek	Bezeq (a company name)	0.09
11	????????		????????	0.15	28	?????????		?????????	0.09
12	הטיסה	ha-tisa	flight the	0.15	29	בתאריך	be-ta'aʁix	on (date)	0.09
13	הביטוח	ha-bituax	insurance the	14	30	אתם	atem	you	0.09
14	המלון	ha-malon	the hotel	0.13	31	נופש	nofeʃ	vacation?	0.09
15	טיסה	tisa	flight	0.13	32	חולים	xolim	patients	0.08
16	אייס	eis	Ace (a company name)	0.13	33	להתנתק	lehitnatek	to be disconnected	0.08
17	מלון	malon	hotel	0.13	34	הממיר	ha-memiʁ	the converter	0.08

Table 17
Top 34 words according to IG for five categories.

#	Word in Hebrew	Trans-literation	translation into English	weight	#	Word in Hebrew	Trans-literation	translation into English	weight
1	רכב	ʁexev	car	0.38	18	טכנאי	texnai	technician	0.12
2	הרכב	ha-ʁexev	the car	0.3	19	במלון	ba-malon	at the hotel	0.11
3	הוט	Hot	Hot (a company name)	0.21	20	סיקסט	sikst	Sixt (a company name)	0.11
4	????		????	0.19	21	חולים	xolim	patients	0.11
5	?????		?????	0.19	22	ההזמנה	ha-hazmana	the booking	0.11
6	??????		??????	0.19	23	לי	li	to me	0.1
7	???		???	0.19	24	נופש	nofeʃ	vacation?	0.1
8	??		??	0.18	25	?????????		?????????	0.1
9	יס	Yes	Yes (a company name)	0.17	26	הממיר	ha-memiʁ	the converter	0.09
10	הטיסה	ha-tisa	flight the	0.16	27	להתנתק	lehitnatek	to be disconnected	0.09
11	????????		????????	0.16	28	בזק	bezek	Bezeq (a company name)	0.09
12	המלון	ha-malon	the hotel	0.15	29	השטיח	ha-ʃatiax	The Carpet (part of a company name)	0.09
13	טיסה	tisa	flight	0.15	30	בתאריך	be-ta'aʁix	on (date)	0.08
14	איסתא	Ista	Issta (a company name)	0.14	31	אותי	oti	me	0.08
15	מלון	malon	hotel	0.14	32	להוט	le-hot	to Hot (a company name)	0.08
16	למלון	la-malon	to the hotel	0.12	33	ברכב	ba-ʁexev	by car	0.08
17	החולים	ha-xolim	the patients	0.12	34	הדקה	ha-daka	The Minute (part of a company name)	0.08

with two suitcases and two handbags … so we walked for about an hour" (category of vacation deals).

From these findings, it can be concluded that the issue that most concerns the people who contacted the companies in question is the quality of the service they provide. It seems that frustration is most evident when the complainants feel that they are not receiving the service they expected to receive. Examples of words from this area that are widely used: talk, ask, disconnect, reach, representative, and service. This was especially evident in the domain of television (more than 60% of the most frequent words belong to this semantic field).

Another issue that significantly preoccupies the complainants is "time" – the time of service provision, the lack of timely service, delays, etc. This was expressed in words such as date, again, years, day, month, minute, hour, and already.

The most common and important issues for the complainants in almost all the company categories is unsatisfactory service combined with late deliveries. These findings are consistent with the research of Rossmann et al. (2017), who found that the quality of service solutions and procedural justice in response to complaints have the strongest impact on customer satisfaction.

A very interesting finding is that in the domain of hospitals only, the subject of the domain itself (e.g., the patient, the medical treatment, and the medical staff) has the greatest importance. This field was expressed in the words: patients, surgery, doctor, nurse, emergency room, examination, treatment and more.

Combining this finding (i.e., the importance of the subject of the complaint) with the finding of various multiple sequences of question marks specifically in the domain of hospitals, can indicate that not only are personal care and human relationships important in this domain, but also this is a very emotional domain in which failure to meet expectations leads to great frustration.

Obviously, companies are motivated primarily by economic considerations. Businesses endeavor to generate customer loyalty primarily for economic reasons. An increase in customer retention has been shown to produce substantial increases in net present value of profits (Reichheld, 1996; Reichheld & Sasser, 1990). According to our analysis, the issue that preoccupies complainants least of all is the price; in almost all categories, it is mentioned least frequently. Much more important to people is good service and a considerate attitude, as well as punctuality. These findings may interest service providers, and help them to improve service, and to focus on the most important fields.

6.2. Analysis of the best accuracy results for six categories

Table 16 in Appendix C presents the top 34 words according to best accuracy result (88.41%) that was achieved by SL with the top 659 word unigrams according to IG for six categories. These words are presented in Hebrew together with their translations into English, their transliterations, and their weights according to IG.

The main findings and conclusions derived from Table 16 are: (1) most of the top occurring words are relevant to specific categories, e.g., car rental, insurance companies, and vacation deals. Additional top words are names of specific companies, which are valuable in identifying the category to which they belong, e.g., Hot, Issta, Bezeq, and Sixt. (2) Most of the words that appear in Table 16 overlap with words that appear in Table 11 (for seven categories). (3) The main difference between the words that appeared in both Tables 11 and 16 is that these words received a higher score in Table 16 (e.g., car [0.31 vs. 0.28]; insurance [0.24 vs. 0.19]; hotel [0.13 vs. 0.11]). A possible explanation is that the corpus discussed in Table 16 is smaller and there are fewer top occurring words, and therefore each such word receives a higher weight. (4) Words that appeared in Table 11 and are not listed in Table 16 are words belonging to the domain of cellular communication, which was filtered out in this experiment. (4) Additional top occurring tokens are pronouns (e.g., “I”, “you”, and “to me”) and sequences of different lengths of question marks (e.g., “??”, “???”, “????”, and “?????”).

6.3. Analysis of the best accuracy result for five categories

Table 17 in Appendix C presents the top 34 words according to best accuracy result (91.21%) achieved by SL using the top 400 word unigrams according to IG for five categories. These top words are presented in Hebrew together with their translations into English, their transliterations, and their weights according to IG.

The main findings and conclusions derived from Table 17 are: (1) Here too, most of the top occurring words are words that are relevant to specific categories. Additional top occurring words are names of specific companies, which are valuable in identifying the category to which they belong. (2) Words that appear in Tables 11 and 16 (seven and six categories, respectively) and are not listed in Table 17 (five categories) are words belonging to the categories of cellular communication and insurance, which were filtered out. (3) In general, many top occurring words that appear in Tables 11 and 16 also appear in Table 17. The main difference is that these words received a higher score in Table 17, e.g., car (0.38 vs. 0.31 and 0.28, respectively); Yes (a company name) (0.17 vs. 0.16 and 0.15), hotel (0.14 vs. 0.13 and 0.11). A possible explanation is that the corpus discussed in Table 17 is smaller and there are fewer top occurring words, and therefore each top occurring word receives a higher weight. (4) Additional top occurring tokens are pronouns (e.g., “me” and “to me”) and sequences of different lengths of question marks (e.g., “??”, “???”, “????”, and “?????”) that were surprisingly helpful.

7. Summary, conclusions and future work

In this study, we present a system that analyzes and classifies complaint letters written in Hebrew according to the company categories such as hospitals, insurance, rental cars, TV, and vacation deals. To the best of our knowledge, this study is one of the first (and perhaps the first) to perform successful classification of complaint letters according to company categories. The classification experiments were performed using various sets of word unigrams, four ML methods, two feature filtering methods, and parameter tuning. The best accuracy results for seven, six, five, and four categories are 84.5%, 88.4%, 91.4%, and 93.8%, respectively.

An analysis of the top occurring content words[3] reveals that what most disturbing to the complainants is receiving unsatisfactory service from the service providers and late deliveries. An interesting finding is that only in the domain of hospitals was the most important issue the subject of the domain itself (i.e., the patient, the medical treatment, the location of the treatment, and the medical staff). Another interesting finding is that the complainants’ and the companies’ attitudes toward “price” seem to be at odds with each other. To the complainants, price is of little or no importance. The service-providing companies, in contrast, see profitability as being of paramount importance. In our opinion, if companies were to give greater importance to service that is personal, fair, and timely (as expressed in our findings, which contain many words from these fields), they would increase the level of customer satisfaction even without giving discounts or financial offers.

Future research proposals include (1) balancing the number of complaint letters for each category in order to prevent bias towards the larger classes; (2) classifying complaint letters according to the severity of their explicit and/or covert verbal violence; (3) investigating the impact of emotions on complaint letters, similar to Rangel and Rosso (2016); (4) conducting additional experiments

[3] Content words are nouns, main verbs, adverbs, and adjectives. It is an open class of words, which plays a crucial role in conveying semantic information (Howell et al., 1999).

for larger corpora of complaint letters according to their severity, written in various languages; (5) implementing other types of features such as character/word n-grams ($n > 1$), relevant key-phrases, and stylistic feature sets (e.g., quantitative features, orthographic features, and part-of-speech [POS] or syntactic features); (6) further and deeper analysis of the question mark series and their interpretation can be done using sentiment analysis tools and methods; (7) applying additional supervised ML methods such as the model suggested by Joulin et al. (2016) on the one hand, and DL methods on the other hand; and (8) building and applying model(s) that will use also keyphrases (HaCohen-Kerner et al., 2007), expansions of abbreviations (HaCohen-Kerner, Kass & Peretz, 2008), and summaries (HaCohen-Kerner, Malin & Chasson, 2003) that can be extracted from the complaint letters.

Acknowledgments

This work was partially funded by the Jerusalem College of Technology (Lev Academic Center) and Bar-Ilan University. The authors gratefully acknowledge the editors and two anonymous reviewers for their fruitful comments and valuable contributions. Thanks also to Rachel Bar-Yosef and Shalom Mashbaum for their English editing.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ipm.2019.102102.

Appendix A

Below, we present several examples taken from letters of complaint from three categories: insurance, communication, and vacation deals.

Insurance

Getting into incidental details:

- "So it's like this – mandatory insurance at 3552 NIS: 20–25% more expensive than the lowest price estimate (there are 3) but let's put that aside."
- "I paid no less than 2639 two months later, in December, when I requested to cancel the insurance and change cars – they refunded only 1461 instead of, in my calculation – (2639/12)*10 = 2200 shekels. Where is my refund money???"

Communication

Overseas calls

- "Overseas calls to numbers overseas did not work at all in the countries where I was."
- "I spent time overseas and in my last bill you charged me 4.90 NIS per minute for 37 min of air-time overseas. This is despite the contract that was signed in my presence that declares that the price is 0.90 NIS per minute."

Overpriced monthly payments

- "It has come to my attention that the monthly payment has rocketed to 1300 NIS."
- "I have discovered that I am being charged 30 NIS more than my obligation to the program."

Vacation deals

Inconsistency between what was promised and the result:

- "There is a significant gap between what we were promised and what we received."
- "We arrived at the hotel that they had told us is 5 stars, but unfortunately we were disappointed."
- "Although it is also 5 stars all-inclusive, it is very inferior."

In addition, it is important to note that some of the complaint letters in Hebrew are written in blunt language, and they can contain overt or covert violence, such as: (1) Curses – "You are thieves, you are cheaters, you are liars – you are really …", (2) Irony – "You reminded us why we do not need outside enemies from outside on such a day!"; and (3) Threats – "Otherwise I will sue you and go to a company that treats its customers with respect!".

Appendix B

While carrying out the research, we noted that many of the most common words that helped us classify the letters according to the company type were associated with the name of the company itself which appeared in the letter; we therefore decided to make two

versions of StopWords. Although both lists are called StopWords, nonetheless, while the first list includes 48 general words in Hebrew (Table 15), the second list includes 238 words that are mostly common content words that are relevant to the fields under examination.

The extended stopword list contains 238 words that are composed of the 48 StopWords included in the basic stopword list and an additional 190 words composed of the names of the related companies, their categories, their nicknames, their initials, and the same words with different prefix combinations in Hebrew.

Appendix C

This appendix contains two tables that present the top 34 words according to IG for six and five categories.

References

Avery, J., Steenburgh, T. J., Deighton, J., & Caravella, M. (2012). Adding bricks to clicks: Predicting the patterns of cross-channel elasticities over time. *Journal of Marketing, 76*(3), 96–111.

Blum, A. L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial intelligence, 97*(1–2), 245–271.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*, 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 26*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.

Brinton, L. J. (2000). *The structure of modern English: A linguistic introduction.* Amsterdam: John Benjamins.

Camilleri, D., & Prescott, T. (2017). Analysing the limitations of deep learning for developmental robotics. *Conference on Biomimetic and Biohybrid Systems* (pp. 86–94). Springer, Cham.

Le Cessie, S., & Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 41*(1), 191–201.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST), 2*(3), 27.

Choe, P., Lehto, M. R., Shin, G. C., & Choi, K. Y. (2013). Semiautomated identification and classification of customer complaints. *Human Factors and Ergonomics in Manufacturing & Service Industries, 23*(2), 149–162.

Conneau, A., Schwenk, H., Barrault, L., & Lecun, Y. (2016). Very deep convolutional networks for text classification. arXiv preprint arXiv:1606.01781.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd international conference on Machine learning* (pp. 233–240). .

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*, 861–874. https://doi.org/10.1016/j.patrec.2005.10.010.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research, 3*, 1289–1305.

Forster, J., & Entrup, B. (2017). A cognitive computing approach for classification of complaints in the insurance industry. *IOP Conference Series: Materials Science and Engineering. 261*IOP Publishing012016.

Fox, G. (2008). Getting good complaining without bad complaining. *Journal of Consumer Satisfaction Dissatisfaction and Complaining Behavior, 21*, 23–40.

Fürnkranz, J. (1998). A study using n-gram features for text categorization. *Austrian Research Institute for Artifical Intelligence, 3*(1998), 1–10.

Gergen, K. J., & Gergen, M. M. (1987). Narratives of relationship. In R. Burnett, P. McGhee, & D. Clarke (Eds.). *Accounting for relationships: Explanation, representation and knowledge* (pp. 267–288). New York, NY, US: Methuen.

Gokulakrishnan, B., Priyanthan, P., Ragavan, T., Prasath, N., & Perera, A. (2012). Opinion mining and sentiment analysis on a twitter data stream. *Advances in ICT for emerging regions (ICTer), 2012 International Conference on* (pp. 182–188). IEEE.

Gonçalves, C. A., Gonçalves, C. T., Camacho, R., & Oliveira, E. C. (2010). The impact of preprocessing on the classification of medline documents. *Proceedings of the 10th International Workshop on Pattern Recognition in Information Systems* (pp. 53–61). .

HaCohen-Kerner, Y., & Blitz, S. Y. (2010). Initial experiments with extraction of stopwords in Hebrew. *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval (KDIR)* (pp. 449–453). October 25-28, 2010. SciTePress 2010, ISBN 978-989-8425-28-7.

HaCohen-Kerner, Y., Dilmon, R., Friedlich, S., & Cohen, D. N. (2016). Classifying true and false Hebrew stories using word N-Grams. *Cybernetics and Systems, 47*(8), 629–649.

HaCohen-Kerner, Y., & Erlich, O. T. (2014). Identifying the correct root of an ambiguous Hebrew word. *Language, Culture, Computation. Computational Linguistics and Linguistics* (pp. 36–53). Springer.

HaCohen-kerner, Y., Ido, Z., & Ya'akobov, R. (2017). Stance classification of tweets using skip char Ngrams. *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)* (pp. 266–278). Springer, Cham.

HaCohen-Kerner, Y., Kass, A., & Peretz, A. (2008). Combined one sense disambiguation of abbreviations. *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers* (pp. 61–64). Association for Computational Linguistics.

HaCohen-Kerner, Y., Kass, A., & Peretz, A. (2010). HAADS: A Hebrew Aramaic abbreviation disambiguation system. *Journal of the American Society for Information Science and Technology, 61*(9), 1923–1932.

HaCohen-Kerner, Y., Malin, E., & Chasson, I. (2003). Summarization of Jewish law articles in Hebrew. *CAINE* (pp. 172–177). .

HaCohen-Kerner, Y., Mughaz, D., Beck, H., & Yehudai, E. (2008). Words as classifiers of documents according to their historical period and the ethnic origin of their authors. *Cybernetics and Systems: An International Journal, 39*(3), 213–228.

HaCohen-Kerner, Y., Stern, I., Korkus, D., & Fredj, E. (2007). Automatic machine learning of keyphrase extraction from short html documents written in Hebrew. *Cybernetics and Systems: An International Journal, 38*(1), 1–21.

Haifeng, X. I. A., & Junhua, C. (2013). Hot complaint intelligent classification based on text mining. *Journal of Shanghai Normal University (Natural Sciences), 42*(5), 470–475 in Chinese.

Hall, M. A. (1998). *Correlation-based Feature Subset Selection for Machine Learning*.

Hall, M. A., & Smith, L. A. (1998). Practical feature subset selection for machine learning. *Computer science'98 proceedings of the 21st Australasian computer science conference ACSC. 98. Computer science'98 proceedings of the 21st Australasian computer science conference ACSC* (pp. 181–191).

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The weka data mining software: An update. *ACM SIGKDD Explorations Newsletter, 11*(1), 10–18.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29–36.

Hayati, S. A., Wicaksono, A. F., & Adriani, M. (2016). Short text classification on complaint documents. *International Journal of Computational Linguistics and Applications, 7*(2), 129–143.

Ho, T. K. (1995). Random decision forests. *Proceedings of the 3rd International Conference on Document Analysis and Recognition* 14–16 August 1995. 278–282.

Howell, P., Au-Yeung, J., & Sackin, S. (1999). Exchange of stuttering from function words to content words with age. *Journal of Speech, Language, and Hearing Research, 42*(2), 345–354.

Hu, X., Tang, J., Gao, H., & Liu, H. (2013, May). Unsupervised sentiment analysis with emotional signals. *Proceedings of the 22nd international conference on World Wide Web* (pp. 607–618). ACM.

Jiang, M., Liang, Y., Feng, X., Fan, X., Pei, Z., Xue, Y., et al. (2018). Text classification based on deep belief network and softmax regression. *Neural Computing and*

Applications, 29(1), 61–70.
Jin, J., Yan, X., Yu, Y., & Li, Y. (2013). Service failure complaints identification in social media: A text classification approach. In: *the Proc. of the Thirty Fourth International Conference on Information Systems* 2013.
Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In: *the Proc. of the 10th European Conference on Machine Learning (ECML)* (pp. 137–142). .
Johnson, R., & Zhang, T. (2016). Convolutional neural networks for text categorization: Shallow word-level vs. deep character-level. arXiv preprint arXiv:1609.00718.
Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.
Kanaris, I., Kanaris, K., Houvardas, I., & Stamatatos, E. (2007). Words versus character n-grams for anti-spam filtering. *International Journal on Artificial Intelligence Tools, 16*(06), 1047–1067.
Kanaris, I., & Stamatatos, E. (2009). Learning to recognize webpage genres. *Information Processing & Management, 45*(5), 499–512.
Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., & Murthy, K. R. K. (2001). Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computation, 13*(3), 637–649. http://doi.org/10.1162/089976601300014493.
Knight, K. (1999). Mining online text. *Communication ACM, 42*(11), 58–61.
Koller, D., & Sahami, M. (1996). *Toward optimal feature selection.* Stanford, CA: Stanford University284–292.
Kwon, O., & Sim, J. M. (2013). Effects of data set features on the performances of classification algorithms. *Expert Systems with Applications, 40*(5), 1847–1857.
Landwehr, N., Hall, M., & Frank, E. (2005). Logistic model trees. *Machine Learning, 59*(1–2), 161–205. http://doi.org/10.1007/s10994-005-0466-3.
Lee, J.Y., & Dernoncourt, F. (2016). Sequential short-text classification with recurrent and convolutional neural networks. arXiv preprint arXiv:1603.03827.
Levin, E. (1990). A recurrent neural network: Limitations and training. *Neural Networks, 3*(6), 641–650.
Liparas, D., HaCohen-Kerner, Y., Moumtzidou, A., Vrochidis, S., & Kompatsiaris, I. (2014). News articles classification using random forests and weighted multimodal features. *Information Retrieval Facility Conference* (pp. 63–75). Springer, Cham.
Mahayudin, N. H. M., Haron, S. A., & Yin-Fah, B. C. (2010). Unpleasant market experience and consumer complaint behavior. *Asian Social Science, 6*(5), 63–69.
Maia, P., Carvalho, R.N., Ladeira, M., Rocha, H., & Mendes, G. (2014). Application of text mining techniques for classification of documents: a study of automation of complaints screening in a Brazilian Federal Agency.
Martins, B., & Silva, M. J. (2005). Language identification in web pages. *Proceedings of the 2005 ACM symposium on Applied computing* (pp. 764–768). ACM.
Mayer, J. E. (1960). Ensembles of maximum entropy. *The Journal of Chemical Physics, 33*(5), 1484–1487.
McCallum, A., & Nigam, K. (1998). A comparison of event models for naive bayes text classification. *AAAI-98 workshop on learning for text categorization. 752. AAAI-98 workshop on learning for text categorization* (pp. 41–48).
Méndez, J. R., Iglesias, E. L., Fdez-Riverola, F., Díaz, F., & Corchado, J. M. (2005). Tokenising, stemming and stopword removal on anti-spam filtering domain. *Conference of the Spanish Association for Artificial Intelligence* (pp. 449–458). Springer.
Mikolov, T., Deoras, A., Povey, D., Burget, L., & Černocký, J. (2011). Strategies for training large scale neural network language models. *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on* (pp. 196–201). IEEE.
Mitchell, T. M., Nowlan, S. J., & Platt, J. C. (1997). *Machine learning*1995*A convolutional neural network hand tracker. advances in neural information processing systems* (1st edition). New York, NY: McGraw-Hill901–908.
Mladenic, D., & Grobelnik, M. (1998). Word sequences as features in text-learning. In: *Proc. of the 17th Electrotechnical and Computer Science Conference (ERK98)* (pp. 145–148). .
Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval, 2*(1–2), 1–135.
Patra, A., & Singh, D. (2013). A survey report on text classification with different term weighing methods and comparison between classification algorithms. *International Journal of Computer Applications, 75*(7).
Pennebaker, J., Matthias, W., Mehl, R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology, 54*(1), 547–577.
Pfeil, M. N., Yersin, B., Trueb, L., Feiner, A. S., & Carron, P. N. (2018). A retrospective study of complaint letters sent to a Swiss emergency department between 2009 and 2014. *Revue d'epidemiologie et de sante publique, 66*, 75–80.
Pinto, M. B., & Mansfield, P. (2012). Facebook as a complaint mechanism: An investigation of millennials. *Journal of Behavioral Studies in Business, 5*, 1–12.
Platt, J. C. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines. *Advances in Kernel Methods, Support Vector Learning, 208*, 1–21 http://doi.org/10.1.1.43.4376.
Pourret, O., Naïm, P., & Marcot, B. (Vol. Eds.), (2008). *Bayesian networks: A practical guide to applications: 73*. Chichester, West Sussex, England: John Wiley & Sons.
Provost, F., Fawcett, T., & Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms. *Proceeding of the 15th International Conference on Machine Learning* (pp. 445–453). Morgan Kaufmann.
Rajaraman, A. & Ullman, J.D. (.2011). “Data mining”. mining of massive datasets (PDF). pp. 1–17. doi:10.1017/CBO9781139058452.002ISBN 978-1-139-05845-2.
Rangel, F., & Rosso, P. (2016). On the impact of emotions on author profiling. *Information processing & management, 52*(1), 73–92.
Reichheld, F. F. (1996). The loyalty effect. 2005 In F. X. Song, S. H. Liu, & J. Y. Yang (Vol. Eds.), *A comparative study on text representation schemes in text categorization. pattern analysis and applications: 8*, (pp. 199–209). Boston, MA: Harvard Business School Press.
Reichheld, F. F., & Sasser, W. E. (1990). Zero defections: Quality comes to services. *Harvard Business Review, 68*(September/October), 105–111.
Rossmann, A., Wilke, T., & Stei, G. (2017). Usage of social media systems in customer service strategies. *Proceedings of the 50th Hawaii International Conference on System Sciences* (pp. 3950–3959). .
Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one, 10*(3), e0118432.
Sasaki, Y. (2007). The truth of the F-measure. *Teach Tutor Material, 1*(5), 1–5.
Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys (CSUR), 34*(1), 1–47.
Secker Z. (1991). Static Elements in the Poetic Text, Distinctive Poetic and Ideological Principles Latent in Secondary Modeling of Words’ Meaning in the Poetry of Yehuda Amichai and his Contemporaries. Doctoral Dissertation. Tel Aviv University. [Hebrew].
Shahana, P. H., & Omman, B. (2015). Evaluation of features on sentimental analysis. *International Conference on Information and Communication Technologies (ICICT 2014). 46. International Conference on Information and Communication Technologies (ICICT 2014)* (pp. 1585–1592). Procedia Computer Science.
Sovran, T. (1994). Meaning, reference, and the semantic fields of abstract notions. *Hebrew Linguistics, 37*, 41–54 Hebrew].
Song, F., Liu, S., & Yang, J. (2005). A comparative study on text representation schemes in text categorization. *Pattern analysis and applications, 8*(1–2), 199–209.
Sumner, M., Frank, E., & Hall, M. (2005). Speeding up logistic model tree induction. *Knowledge Discovery in Databases: PKDD 2005. 3721. Knowledge Discovery in Databases: PKDD 2005* (pp. 675–683). Springer. http://doi.org/10.1007/11564126.
Sun, W., Tseng, T. L. B., Zhang, J., & Qian, W. (2017). Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data. *Computerized Medical Imaging and Graphics, 57*, 4–9.
Surjandari, I., Megawati, C., Dhini, A., & Hardaya, I. S. (2016). Application of text mining for classification of textual reports: A study of indonesia's national complaint handling system. *Proceedings of the 2016 International Conference on Industrial Engineering and Operations Management Kuala Lumpur* (pp. 1147–1156). March 8-10, 2016.
Swain, P. H., & Hauska, H. (1977). The decision tree classifier: Design and potential. *IEEE Transactions on Geoscience Electronics, 15*(3), 142–147.
Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics, 37*(2), 267–307.
Tax, S., Brown, S., & Chandrashekaran, M. (1998). Customer evaluations of service complaint experiences: Implications for relationship marketing. *Journal of Marketing, 62*, 60–76.
Tripp, T. M., & Gregoire, Y. (2011). When unhappy customers strike back on the internet. *Sloan Management Review, 52*(3), 37–44.
Weinberger, K., Dasgupta, A., Attenberg, J., Langford, J., & Smola, A. (2009). Feature hashing for large scale multitask learning. arXiv preprint arXiv:0902.2206.
Whitney, A. W., & Dwyer, S. J. (1966). Performance and implementation of the k-nearest neighbor decision rule with incorrectly identified training samples. In: *Proc.*

4th Annu. Allerton Conf. Circuit and System Theory (pp. 96–106). .
Wintner, S. (2004). Hebrew computational linguistics: Past and future. *Artificial intelligence review, 21*(2), 113–138.
Witten, I. H., & Frank, E. (2005). Data mining: Practical machine learning tools and techniques. *Morgan Kaufmann*. San Francisco.
Yamasaki, T., Fukushima, Y., Furuta, R., Sun, L., Aizawa, K., & Bollegala, D. (2015, October). Prediction of user ratings of oral presentations using label relations. *Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia* (pp. 33–38). ACM.
Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. *In Icml, 97*, 412–420.