Distinguishing between True and False Stories using various Linguistic Features

Yaakov HaCohen-Kerner
Dept. of Computer Science
Jerusalem College of Technology
21 Havaad Haleumi St., P.O.B. 16031
9116001 Jerusalem, Israel
kerner@jct.ac.il

Rakefet Dilmon
Dept. of Hebrew
and Semitic Languages
Bar-Ilan University
5290002 Ramat-Gan, Israel
rak2@bezeqint.net

Shimon Friedlich
Dept. of Computer Science
Jerusalem College of Technology
21 Havaad Haleumi St., P.O.B. 16031
9116001 Jerusalem, Israel
shimonfriedlich@gmail.com

Daniel Nissim Cohen
Dept. of Computer Science
Jerusalem College of Technology
21 Havaad Haleumi St., P.O.B. 16031
9116001 Jerusalem, Israel
sdanielco@gmail.com

Abstract

This paper analyzes what linguistic features differentiate true and false stories written in Hebrew. To do so, we have defined four feature sets containing 145 features: POS-tags, quantitative, repetition, and special expressions. The examined corpus contains stories that were composed by 48 native Hebrew speakers who were asked to tell both false and true stories. Classification experiments on all possible combinations of these four feature sets using five supervised machine learning methods have been applied. The Part of Speech (POS) set was superior to all others and has been found as a key component. The best accuracy result (89.6%) has been achieved by a combination of sixteen POS-tags and one quantitative feature.

1 Introduction

"A lie is a false statement to a person or group made by another person or group who knows it is not the whole truth, intentionally" (Freitas-Magalhães, 2013). Dilmon (2014) defines a lie as "a linguistic message that conveys a falsehood or in which the truth is intentionally manipulated, in order to arouse in the listener a belief which he would not otherwise have held."

The efforts to discover linguistic cues to detect lies are based on the assumption that there are differences between the language of an individual when he (or she) is not telling the truth and his (or her) “normal,” truthful language. Fraser (1991) claims that these differences are the outcome of a feeling of stress, which is manifest in a decline in capacity for cognitive integration, in precision, in organization, and in ranking things. These difficulties result in a change in the normal elements of the speaker’s language.

There were a few studies during the last four decades concerning verbal cues that characterize a lie discourse. Dulaney (1982) finds that the response time was shorter, there were fewer special words, a smaller number of verbs in the past tense, and a faster speech rhythm when an individual was lying; there were fewer words in the discourse, as well as a tendency to short messages. Knapp et al. (1974) find that there were more general declarations and fewer factual ones, linguistic ambiguity, repeated declarations, more markers of diminishment (few, a little, hardly) and fewer group markers (we, our, all of us), more markers of the other (they) and fewer personal declarations (I, me). Hollien and Rosenberg (1991) use lexical breakdown to investigate deception (type-token ratio - TTR), and finds less linguistic diversity when a person is practicing deception.

29th Pacific Asia Conference on Language, Information and Computation: Posters, pages 176 - 186
Shanghai, China, October 30 - November 1, 2015

The studies of Dilmon (2007; 2008; 2012) conduct a comprehensive examination of the linguistic criteria that differentiate between the discourse of truth and of deception in the Hebrew language, and attempt to produce a primary test of the cognitive and emotional functions involved in the latter type of discourse. Forty three verbal criteria (Section 2.2) were classified according to the cognitive and emotional functions affecting the speaker, also addressing his level of awareness of these functions. Except one verbal criterion that was automatically computed by a program, the values of all other criteria for each story were computed by hand. This study starts from the end of the studies of Dilmon. Firstly, we implemented and/or applied four feature sets: POS-tag features, quantitative features, repetition features, and special expressions. Secondly, the application of the features is automatically done by a computer program in contrast to Dilmon's features (42 of her 43 features were computed by hand for each story). Thirdly, in contrast to Dilmon's studies that found which are the specific criteria that are statistically significant differentiators, we apply five supervised machine learning (ML) methods and various combinations of feature sets to find the best method for single-document classification, i.e., for each input story identifying whether it is a true or a false story. That will potentially lead to find discoveries concerning distinguishing between truth and false stories.

The task of distinguishing between true and false story as well as the interpretation of the obtained results are of practical interest for any language in general and for Hebrew in particular. Such a system can be of great help to the work of organizations, such as workplaces, detective agencies, police, and courts, to identify various types of stories.

The rest of this paper is organized as follows: Section 2 presents relevant background on linguistic examination in relevant systems, linguistic examination between discourses of truth and deception, text classification, and text classification of deception and true stories. Section 3 describes the classification model and the chosen feature sets. Section 4 presents the examined corpus, the experimental results and their analysis. Finally, Section 5 summarizes the main findings and suggests future directions.

2 Relevant Background

2.1 Linguistic examination in relevant systems

Argamon et al. (2009) describe an automatic process that profiles the author of an anonymous text. Accurate profiling of an unknown author is important for various tasks such as criminal investigations, market research, and national security. The deciphering the profile of someone is performed in the following way: Given a corpus of documents, marked as "male" and "female". Only four features were selected: sex, age, mother tongue, and neurotic level of disturbance behavior. Combination of linguistic features and various ML methods (Support vector machines and Bayesian regression) enable an automated system to effectively determine several such aspects of an anonymous author.

Chaski (2005) presents a computational, stylometric method that has obtained 95% accuracy and has been successfully used in investigating and adjudicating several crimes involving digital evidence. Chaski's approach focuses on language features that are easily achievable, e.g., word length, sentence length, word frequency, and the distribution of words according to different lengths.

Strous et al. (2009) describe an automatic process that characterizes and identifies schizophrenia in writing. This study investigates and analyzes computer texts written by 36 schizophrenia patients. Each document contains from 300 to 500 words. The system tested differences between these documents to documents written by people who are not sick with this disease. Observations have shown that methods using lexical and syntactic features obtained 83.3% accuracy. 60 features were chosen for the classification process: the 25 most frequent words in the corpus, the 20 most frequent letter tri-grams, and the average number of 15 repetitive words. The main conclusions are: (1) Some of the basic processes in schizophrenia are evident in writing; (2) Automatically identified characteristics of schizophrenic writing are closely related to the clinical description of the disorder; and (3) Automatic classification of samples in writing of schizophrenia is possible.

2.2 Linguistic examination between discourses of truth and deception

Hancock et al. (2005) found that “liars tended to produce more words, fewer first person singular but more third person pronouns, and more sense words than truth-tellers”. Only a small number of criteria were examined, the discourse being studied was written on a computer, the motivation to lie came from preliminary instructions, and the discourse examined was a conversation (not a full text).

The studies of Dilmon (2007; 2008; 2012) dealt with discovering linguistic differences between the discourse of truth and discourse of deception. Dilmon's studies present an investigation of 48 couples of stories told by 48 subjects. Each of them told a true story and a false story. The comparison was made using linguistic instruments, and the results obtained were examined statistically. The 48 subjects are native Hebrew speakers of both sexes, of different ages and a variety of backgrounds (with no criminal background).

The subjects were being instructed to take part in a game in which they had to tell two stories from their past, one true and the other an invention, and the “real” subjects would have to guess which of the stories was true and which an invention. In this way, the subjects themselves chose where and how they would mislead, and they would be motivated to provide stories that would make it hard to identify them as stories of deception. That is to say, they tried to escape detection, as would be the case in an actual deceptive situation. Apart from this instruction, they received no other instructions as to subject matter, length, or any other issue of the story’s substance.

Dilmon (2012) compared between the true stories and the false stories. Her assumption is that the true stories indicate the subject’s ordinary, “normal” language, while the false stories indicate deviations from that normal language. 43 criteria were defined by her to analyze the language of truth and falsity. Part of the criteria were translated to Hebrew from the foreign literature. Other criteria were collected after interviews with an attorney, a police investigator, a military police investigator, and two psychologists who had worked for the police. These criteria belong to the following areas: morphology, syntax, semantics, discourse analysis, and speech prosody.

42 out of 43 criteria were calculated manually. All these criteria were examined whether they differentiate between the discourse of truth and of deception. Statistical analyses using MANOVA were performed with repeated measures for each linguistic criterion. 19 criteria were found to differentiate significantly between the two types of discourse. 5 out of the 19 criteria that have been found as significant belong to the morphology area as follows: 1- # of past tense verbs, 2- # of present tense verbs, 3- # of future tense verbs, 4- # of first person verbs, and 5- # of third person verbs. All these criteria are normalized by the # of verbs in the tested story.

2.3 Text classification

Text classification (TC) is a supervised learning task that assigns natural language text documents to one or more predefined categories (Sebastiani, 2002). The TC task is one of the most fundamental tasks in data mining (DM) and machine learning (ML) literature (Aggarwal and Zhai, 2012).

TC has been applied in various domains, e.g., document indexing, document filtering, information retrieval (IR), information extraction (IE), spam filtering, text filtering, text mining, and word sense disambiguation (WSD) (Pazienza, 1997; Knight, 1999; Sebastiani, 2005).

There are two main types of TC: TC according to categories and to stylistic classification. TC according to categories (e.g., disciplines, domains, and topics) is usually based on content words and/or n-grams (Cavnar and Trenkle, 1994; Damashek, 1995; Martins and Silva, 2005; Liparas et al., 2014).

Literature documents, for instance, are different from scientific documents in their content words and n-grams. However, stylistic classification, e.g., authorship attribution (Stamatatos, 2009; Koppel et al., 2011), ethnicity/time/place (HaCohen-Kerner et al., 2010A; 2010B), genre (Stamatatos, 2000; Lim et al., 2005), gender (Hota et al., 2006; Koppel et al., 2002), opinion mining (Dave et al., 2003), computer science conference classification (HaCohen-Kerner et al., 2013), and sentiment analysis (Pang et al., 2002), is usually based on various linguistic features, such as function words, orthographic features, parts of speech (POS) (or syntactic) features, quantitative features, topographic features, and vocabulary richness.

2.4 TC of deception and true stories

Mihalcea and Strapparava (2009) present initial experiments in the recognition of deceptive language. They introduce three data sets of true and lying texts containing 100 true and 100 false statements for each dataset. They use two classifiers: Naïve Bayes and SVM. Their features were words belonging to several special word classes, e.g., friends (friend, companion, body), and self (our, myself, mine, ours). No feature selection was performed, and stopwords were not removed. Using a 10-fold cross-validation test their accuracy results were around 70%.

Ott et al. (2011) develop a dataset containing 400 truthful hotel reviews and 400 deceptive hotel reviews. Their features were Linguistic Inquiry and Word Count (LIWC) features extracted by the LIWC software (Pennebaker et al., 2007), relative POS frequencies extracted by the Stanford Parser (Klein and Manning, 2003) and 3 n-gram feature sets (unigrams, bigrams, and trigrams). Ott et al. show that the detection of deceptive opinion spam is well beyond the capabilities of human judges. Using Naïve Bayes and SVMlight (Joachims, 1999) and a 5-fold cross-validation test they have found that a bigram-based classification based on unigrams and bigrams obtained an accuracy of 89.6%, and a combination of LIWC features, unigrams and bigrams performed slightly better (89.8%).

3 The Classification Model and the Chosen Feature Sets

We decided to use Dilmon's stories as our data set. We defined, programmed and automatically calculated features for the input stories. In contrast to Dilmon, who calculated the ability of each feature alone to statistically distinguish between true and false stories, we investigated the ability of various combinations of features to classify between true and false stories using various ML methods.

The main stages of the model are as follows:

1. Building a corpus containing 96 stories (48 false and 48 true stories).
2. Computing all four feature sets including the POS-tag features using the tagger built by Adler (Adler, 2007; Adler et al., 2008). This tagger achieved 93% accuracy for word segmentation and POS tagging when tested on a corpus of 90K tokens.
3. Applying five ML methods for each possible combination of feature sets using default parameters.
4. Filtering out non-relevant features using InfoGain (IG) (Yang and Pedersen, 1997) and re-applying the best ML method found in stage #3.

Features

In this paper, we consider 145 features divided into four meaningful linguistic feature sets as follows: 123 POS-tag features, 4 quantitative features, 9 repetition features, and 9 special expressions. These four feature sets have neither been defined nor applied by Dilmon. In this research, some of Dilmon's (2008) criteria have not been examined (e.g., discourse analysis and prosodic elements as stuttering and hesitation marks) because it was difficult to automatically detect them. However, features such as tense verbs and person verbs have been applied among the POS-tag feature set.

We did not choose the bag of words (BOW) or N-gram (which are usually the most frequent continuous sequences of N-grams) as features because they are too simple; they have less meaning and they can be partially seen as a black box. As an example of their low significance is the fact that the linear ordering of the N-grams within the text is ignored. That is to say, these representations are essentially independent of the sequence of words in the collection.

The first chosen feature set contains 123 POS-tag features automatically extracted by Adler's tagger for the Hebrew language (Adler, 2007; Adler et al., 2008). This set contains features, which belong to many feature sub-sets: 7 prefix types, 28 part-of-speech tags, 3 gender types, 5 number types, 4 person types, 3 status types, 7 tense types, 4 pronoun types, 8 named-entity types, 4 interrogative types, 3 prefix types, 15 punctuation types, 5 number suffix types, 4 person suffix types, 2 polarity types, 7 Hebrew verbal stem types, 3 conjunction types, 5 number types, 3 gender suffix types, and 3 quantifier types.

The second feature set is the quantitative set containing 4 types of average # of letters per word, average # of letters per sentence, average # of words per sentence, and TTR (the number of different word types in a text divided by the total number of word tokens).

The third feature set is the repetition features containing the following 9 features: normalized # of n-gram words (for n=1, 2, 3, 4) that repeat themselves in the same sentence, respectively, normalized # of 'ha' (i.e., "the", the definite article in Hebrew), and normalized # of n-gram words (for n=1, 2, 3, 4) that repeat themselves in the entire text only once, twice, 3, or 4 times, respectively. The normalization is done by a division of the computed value to the number of word tokens in the document.

The fourth and the last feature set is the special expressions set that contains the normalized # of the following 9 features: intensifiers, minimizing markers, negative expressions, positive expressions, time expressions, expressions of doubt, Emotive words and words describing emotions, demonstrative pronouns, generalized words, 'et' (a term used to indicate a direct object), and 'shel' (i.e., of, belonging to).

4 Corpus and Experimental Results

The examined corpus (supplied by Dilmon) contains 96 stories (48 false and 48 true stories) that were told by 48 native Hebrew speakers (23 men and 25 women) between the ages of 20 and 45. The reasons for relatively small number of subjects are: (1) The subjects did not receive payment for their participation; each one of them volunteered to participate. It is not easy to find many volunteers for such action. (2) The course of Dilmon's study included a recording of the stories, varying in length from five minutes to an hour. Then an accurate transcription of the stories was required (receiving over 100 pages of transcribed text) and a careful count of all the linguistic characteristics. Table 1 presents general information about this corpus.

Type of story	Total # of words	Avg. # of words per story	Median value of words per story	Std. of words per story
True	8722	181.7	155.5	145.03
False	6720	140	113.5	103.09

Table 1. General information about the corpus.

Five supervised ML methods including two decision tree methods have been selected. The accuracy rate of each ML method was estimated by a 10-fold cross-validation test. These ML methods include SMO and Naïve Bayes (that were examined in the two previous studies about true/false classification mentioned in sub-section 2.4). The five applied ML methods are:

(1) Reduced Error Pruning (REP)-Tree is a fast decision tree learner, which builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning with back fitting (Witten and Frank, 2005). This algorithm sorts values for only numeric attributes. Missing values are dealt with by splitting the corresponding instances into pieces. Because the tree grows linearly with the size of the samples presented, and that, after a while, no accuracy is gained through the increased tree complexity, pruning becomes helpful if used carefully (Elomaa and Kääriäinen, 2001).

(2) J48 is an improved variant of the C4.5 decision tree induction (Quinlan, 1993; Quinlan, 2014) implemented in WEKA. J48 is a classifier that generates pruned or unpruned C4.5 decision trees. The algorithm uses greedy techniques and is a variant of ID3, which determines at each step the most predictive attribute, and splits a node based on this attribute. J48 attempts to account for noise and missing data. It also deals with numeric attributes by determining where thresholds for decision splits should be placed. The main parameters that can be set for this algorithm are the confidence threshold, the minimum number of instances per leaf and the number of folds for REP. As described earlier, trees are one of the easiest thing that could be understood because of their nature.

(3) Sequential Minimal Optimization (SMO; Platt 1998; Keerthi et al. 2001; Hastie and Tibshirani, 1998) is a variant of the Support Vectors Machines (SVM) ML method (Cortes and Vapnik 1995; Vapnik 2013). The SMO technique is an iterative algorithm created to solve the optimization problem often seen in SVM techniques. SMO divides this problem into a series of smallest possible sub-problems, which are then resolved analytically.

(4) Logistic regression (LR; Cessie et al., 1992) is a variant of a probabilistic statistical classification model that is used for predicting the outcome of a categorical dependent variable (i.e., a class label) based on one feature or more (Cessie et al., 1992; Landwehr et al., 2005; Sumner et al., 2005).

(5) Naïve Bayes (NB; John and Langley, 1995; McCallum and Nigam, 1998) is a set of probabilistic classifiers with strong (naive) independence assumptions between the features. The Naive Bayes Classifier method is usually based on the so-called Bayesian theorem (the current probability is computed based on a previous related probability) and is particularly suited when the number of the features is high.

These ML methods have been applied using the WEKA platform (Witten and Frank, 2005; Hall et al., 2009) using the default parameters. After finding the best ML method we have performed further experiments using only this method. Non-relevant features were filtered out using Information gain (InfoGain, IG), a feature selection metric for text classification. IG is a popular measure of feature goodness in text classification (Yang and Pedersen, 1997). It measures the number of bits of information obtained for category prediction by knowing the presence or absence of a feature. In their comparative study, Yang and Pedersen reported that IG and Chi performed best in their multi-class benchmarks. Forman (2003) reported that IG is the best filtering method when one is limited to 20-50 features. In Forman's experiments, IG dominates the performance of Chi for every size of the feature set. The accuracy of each ML method was estimated by a 10-fold cross-validation test.

Combinations of feature sets	Rep-Tree	J48	SMO	LR	NB
P	60.4	68.8	**80.2**	63.5	78.1
Q	61.5	61.5	67.7	64.6	66.7
R	52.1	57.3	60.4	61.5	61.5
S	68.8	66.7	77.1	77.1	68.8
P, Q	62.5	66.7	82.3	68.8	79.2
P, R	60.4	70.8	75.0	65.5	78.2
P, S	57.3	67.7	**83.3**	62.5	79.2
Q, R	60.4	67.7	63.5	67.7	67.7
Q, S	70.8	69.8	77.1	76.0	74.0
R, S	70.8	60.4	75.0	75.0	68.8
P, Q, R	62.5	67.7	77.1	67.7	79.2
P, R, S	71.9	65.6	81.3	71.9	79.2
P, Q, S	58.3	66.7	**84.4**	76.0	81.3
Q, R, S	68.8	66.7	75.0	76.0	70.8
P, Q, R, S	58.3	63.5	81.3	69.8	80.2

Table 2. Accuracy results for the classification of True/False stories.

In this research, there are four feature sets (section 3): POS-tags (P), Quantitative (Q), Repetitions (R), and Special Expressions (S). Therefore, there are 24 = 16 combinations of feature sets (including the empty set). For each ML method we tried all 15 non-empty combinations of feature sets.

Table 2 presents the accuracy results for the classification of true/false stories according to all 15 combinations of feature sets. These results were obtained by applying the 5 supervised ML methods mentioned in Section 3.

Several general conclusions can be drawn from Table 2:

- The first 4 rows present the accuracy results using only one feature set. The best result for 3 ML methods (SMO, J48 and NB) was achieved by the POS-tags set. The best result out of these results was obtained by the POS-tags set using SMO. Similar to Ott et al. (2011) we related to the accuracy results achieved by the POS-tag features (80.2%) as the baseline with which to compare our other results.
- The POS-tags feature set (80.2%) is superior to the other single sets. Several possible explanations for this finding are: this set includes the largest number of features (123), and these features include widespread information about the whole text, which is relevant to the task at hand.
- The SMO method obtained the best accuracy result results for most of the set combinations (in 8 out of 15 experiments).
- The best accuracy result using a combination of 2 sets (83.3%) was obtained using a combination of the POS-tags and the special expressions.
- The best accuracy result in Table 2 (84.4%) was obtained using a combination of 3 sets: POS-tags, quantitative and the special expressions.
- The addition of the repetitions features to the 3 sets (i.e., the combination of all 4 sets) led to a decline in the results (81.3%). The repetitions set was the set with the worst results compared with the other sets for all five ML methods.
- The improvement rate from the best set to the best combination of sets is 4.2%.

Since SMO has been found as the best ML method for our classification task, we decided to do further experiments using only SMO and IG (as explained above).

Combinations of feature sets	SMO before IG		SMO after IG							
	# of feat.	Acc.	# of feat.	Acc.	True			False		
					P	R	F	P	R	F
P	123	80.2	15	86.5	80.9	79.2	80.0	79.6	81.3	80.4
Q	4	67.7	1	57.3	68.4	27.1	38.8	54.5	87.5	67.2
R	9	60.4	1	53.1	53.1	54.2	53.6	53.2	52.1	52.6
S	9	77.1	4	69.8	67.3	77.1	71.8	73.2	62.5	67.4
P, Q	127	82.3	17	**89.6**	91.3	87.5	89.4	88.0	91.7	89.8
P, R	132	75.0	17	87.5	90.9	83.3	87.0	84.6	91.7	88.0
P, S	132	83.3	20	85.4	88.6	81.3	84.8	82.7	89.6	86.0
Q, R	13	63.5	2	63.5	65.1	58.3	61.5	62.3	68.8	65.3
Q, S	14	77.1	5	78.1	77.6	79.2	78.4	78.7	77.1	77.9
R, S	13	75.0	4	69.8	67.3	77.1	71.8	73.2	62.5	67.4
P, Q, R	136	77.1	18	88.5	91.1	85.4	88.2	86.3	91.7	88.9
P, R, S	136	81.3	20	85.4	88.6	81.3	84.8	82.7	89.6	86.0
P, Q, S	137	84.4	21	89.6	93.2	85.4	89.1	86.5	93.8	90.0
Q, R, S	22	75.0	6	79.2	78.0	81.3	79.6	80.4	77.1	78.7
P, Q, R, S	145	81.3	22	89.6	93.2	85.4	89.1	86.5	93.8	90.0

Table 3. Accuracy results for combinations of feature sets using SMO and IG.

Table 3 presents the accuracy results for all combinations of feature sets using SMO (the best ML method according to Table 2) before and after filtering out non-relevant features using IG. In addition, for the stage after activating IG we also present the precision, recall, and F-score results for each type of story (true, false) for all possible combinations of the four feature sets. The following conclusions can be drawn from Table 3 regarding the classification of True/False stories using SMO and IG:

- The best accuracy result (89.6%) has been achieved by three different combination sets. The combination with the smallest number of feature sets, is the combination of two sets: POS-tag and quantitative, which contains 17 features including 16 POS-tag features and one quantitative feature.
- The improvement rate of this combination of two sets from the initial state before performing IG to the state after performing IG is 7.3%. This improvement has been achieved due to the filtering out of 110 features out 127!
- The relatively similar accuracy, precision, recall, and F-score results for both types of stories (true, false) for all types of set combinations represent that the classification results are at the same level of quality for both types of stories.
- By looking at the results of the best combinations in Table 3 (colored with red and blue), we see that on the one hand, the precision values are higher for the true stories (i.e., less false positives; which means that the system has a high ability to present only relevant true stories), and on the other hand, the recall values are higher for the false stories (i.e., less false negatives; which means that the system has a high ability to present all relevant false stories)

Detailed results for the best combination (16 POS-tag features and one quantitative feature) using SMO and IG are presented in Tables 4 and 5. Table 4 presents the suitable confusion matrix and Table 5 shows the values of the ROC and PRC areas. The area under the ROC curve (Bradley 1997; Fawcett 2006) and the area under the PRC curve, i.e., the area under the precision-recall curve (Boyd et al., 2013) are often used to evaluate the performance of ML methods.

		Actual answer	
		True	False
Classifier's answer	True	TP=44	FP=4
	False	FN=6	TN=42

Table 4. The confusion matrix.

	True	False
ROC area	89.6	89.6
PRC area	86.1	84.8

Table 5. The ROC and PRC areas.

Using the TP, FP, FN, FP, and TN values in the confusion matrix, are computed the four popular measures: recall, precision, accuracy and f-measure (Table 3). The ROC area is around 90% and the PRC area is around 85%-86% indicating very good classification performance of the SMO method using the 17 chosen features.

Another deeper observation shows several interesting findings about the most distinguishing features according to the IG method (i.e., features that received the highest weights). Table 6 presents some distinguishing POS features.

Distingin. POS features	**Finding**	**Meaning**
Person-1 (first person)	The average of this feature for the true stories is significantly higher	Truthful people use relatively more first person pronouns
Person-3 (third person)	The average of this feature for the false stories is significantly higher	Liars use relatively more third person pronouns
POS-negation (negation words)	The average of this feature for the false stories is significantly higher	Liars use relatively more negation words

Table 6. Distinguishing POS features according to SMO and IG.

Our findings concerning the use of first-person pronouns, and negative words are consistent with the conclusions of Hancock et al. (2005) who found that the discourse of deception used fewer first-person pronouns, and more negative words. Our findings concerning use of first and third person pronouns are also consistent with the conclusions of Knapp et al. (1974) who found that a lie discourse contains more markers of the other and fewer personal declarations (I, me).

Furthermore, our findings are also consistent with some of Dilmon (2008): (1) The use of negative words in the false stories might reveals the speaker's negative attitude toward his invention, and his insecurity from being in the position of misleading the listener, and (2) Higher use of verbs in the third person and minimal use of verbs in the first person in false stories may imply the speaker's desire to distance himself from a description of the event and from the possibility of accepting responsibility for his actions.

From a pragmatic standpoint, a deception is a deviation from Grice's (1975) "Cooperative Principle", which is subdivided into 4 maxims: of quantity, of quality, of relation, and of manner. He stresses that the meticulous observance of the maxim of quality is a fundamental pre-condition that ensure the operation of the other maxims. Mey (2001) claims that concealment technics (e.g., deliberate omission, and uninformative or disinformative remarks) contradict the Cooperative Principle of Grice. By using negative words and third person verbs, the speaker is violating the maxim of quality.

5 Summary and Future Work

In this paper, we present a methodology for distinguishing between true and false stories based on various linguistic features. The POS-tag set containing 123 features was superior to all other sets with an accuracy result of 80.2%. The best accuracy result (89.6%) was obtained by SMO and IG using two feature sets including only sixteen POS-tag features and one quantitative feature. These results suggest that stylistic differences between any types of true and false stories can be quantified along the lines presented in this paper.

The main contribution of this research is the careful feature set engineering based on analyses construction of feature sets derived from previous studies. This together with the competition between five well-known supervised ML methods, and filtering out of non-relevant features using IG for SMO (the best found ML method), yields considerably improved accuracy results.

Future research proposals are: (1) Apply this classification model to other types of true and false stories coming from other domains and written in various languages; (2) Implement feature sets with a focus on special compound linguistic features that differentiate between true and false stories, speech features such as hesitations or repetitions, n-gram features and other types of stylistic feature sets; (3) Perform experiments to see if some interactions at feature level, not feature set level, have any impact on the classification accuracy; and (4) Perform experiments of distinguishing between true and false stories by people, and comparing their results versus those performed by our system.

References

Meni Adler. 2007. Hebrew Morphological Disambiguation: An Unsupervised Stochastic Word-based Approach, Ph.D. Dissertation, Ben Gurion University, Israel.

Meni Adler, Yael Netzer, David Gabay, Yoav Goldberg, and Michael Elhadad. 2008. Tagging a Hebrew Corpus: The Case of Participles, In *Proceedings of the* LREC-2008, European Language Resources Association, Marrakech, Morocco.

Charu C. Aggarwal and ChengXiang Zhai. 2012. Mining text data. New York, NY: Springer.

Kendrick Boyd, Kevin H. Eng, and C. David Page. 2013. Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals. In *Proceedings of the* Machine learning and knowledge discovery in databases, pages 451-466. Springer Berlin Heidelberg.

Andrew P. Bradley. 1997. The use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms. Pattern Recognition 30: 1145–1159. doi: 10.1016/S0031-3203(96)00142-2

Leo Breiman. 2001. Random Forests. Machine Learning, 45(1): 5–32.

William B. Cavnar and John M. Trenkle. 1994. N-Gram-Based Text Categorization. Ann Arbor MI, 48113(2): 161-175.

Renato F. Corrêa and Teresa B. Ludermir. 2002. Automatic Text Categorization: Case Study, In *Proceedings of the VII Brazilian Symposium on Neural Networks*, SBRN 2002, page 150, IEEE

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector Networks. *Machine learning*, 20(3): 273–297.

Marc Damashek. 1995. Gauging Similarity with N-grams: Language-independent Categorization of Text, Science, 267(5199): 843-848.

Kushal Dave, Steve Lawrence, and David M. Pennock. 2003. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. In *Proceedings of the 12th international conference on World Wide Web*, pages 519-528, ACM.

Rakefet Dilmon. 2004. Linguistic Differences between Lie and Truth in Spoken Hebrew – Doctoral Dissertation. Bar Ilan University, Ramat Gan, Israel (in Hebrew).

Rakefet Dilmon. 2007. Fiction or Fact? Comparing True and Untrue Anecdotes, *Hebrew Linguistics*, 59: 23-42 (in Hebrew).

Rakefet Dilmon. 2008. Between Thinking and Speaking - Linguistic Tools for Detecting a Fabrication, *Journal of Pragmatics*, 41(6): 1152-1170.

Rakefet Dilmon. 2012. Linguistic Examination of Police Testimony – Falsehood or Truth? In: R. Peled-Laskov, E. Shoham & M. Carmon (eds.), False Convictions: Philosophical, Organizational and Psychological Aspects (433 pages), Perlstein-Ginosar & Ashkelon Academic College, Tel-Aviv, pages 95-114 (in Hebrew).

Rakefet Dilmon. 2013. False speech, linguistic aspects in Hebrew, in: D. A. Russell, G. Khan, & D. L. Vanderzwaag (eds.), The Encyclopedia of Hebrew Language and Linguistics, Leiden, Brill, Boston and Tokyo, pages 542-546.

Earl F. Dulaney. 1982. Changes in Language Behavior as a Function of Veracity. Human Communication Research, 9(1): 75-82.

Tapio Elomaa, Matti Kääriäinen. 2001. An Analysis of Reduced Error Pruning. *Journal of Artificial Intelligence Research* 15: 163–187. doi: 10.1613/jair.816

Tom Fawcett, 2006. An Introduction to ROC Analysis. *Pattern Recognition Letters*, 27(8): 861–874. doi: 10.1016/j.patrec.2005.10.010

George Forman. 2003. An Extensive Empirical Study of Feature Selection Metrics for Text Classification, The *Journal of machine learning research*, 3: 1289-1305.

Bruce Fraser. 1991. Questions of Witness Credibility. Working Papers Series, Program on Negotiation. Cambridge, MA: Harvard Law School, pages 3-91.

Armindo Freitas-Magalhães. 2013. The Face of Lies. Porto: FEELab Science Books. ISBN 978-989-98524-0-2.

Herbert P. Grice. 1975. Logic and conversation. In Cole Peter & Jerry L. Morgan (Eds.), *Syntax and semantics*, vol. 3 Speech acts: 41-58. New York: Academic Press.

Yaakov HaCohen-Kerner, Hananya Beck, Elchai Yehudai, and Dror Mughaz. 2010A. Stylistic Feature Sets as Classifiers of Documents According to their Historical Period and Ethnic Origin. *Applied Artificial Intelligence*, 24(9): 847-862.

Yaakov HaCohen-Kerner, Hananya Beck, Elchai Yehudai, Mordechay Rosenstein, and Dror Mughaz. 2010B. Cuisine: Classification using Stylistic Feature Sets and/or Name-Based Feature Sets. *Journal of the*

American Society for information Science & Technology (JASIST), 61(8): 1644-1657.

Yaakov HaCohen-Kerner, Avi Rosenfeld, Maor Tzidkani, and Daniel Nisim Cohen. 2013. Classifying Papers from Different Computer Science Conferences. In *Proceedings of the Advanced Data Mining & Applications*. ADMA 2013, Part I, LNAI 8346, pages 529-541, Springer Berlin Heidelberg.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software. *ACM SIGKDD Explorations Newsletter*, 11(1):10.

Jeffrey T. Hancock, Lauren Curry, and Saurabh Goorha, Michael Woodworth. 2005. Automated Linguistic Analysis of Deceptive and Truthful Synchronous Computer-Mediated Communication. In *Proceedings of the 38th Hawaii International Conference on System Science*, pages 1-10.

Trevor Hastie and Robert Tibshirani. 1998. Classification by Pairwise Coupling. *The annals of statistics*, 26 (2): 451-471.

Harry Hollien and Aaron E. Rosenberg. 1991. The Acoustics of Crime: The new Science of Forensic Phonetics. New York: Plenum.

Sobhan R. Hota, Shlomo Argamon, and Rebecca Chung. 2006. Gender in Shakespeare: Automatic Stylistics Gender Character Classification Using Syntactic, Lexical and Lemma Features. Digital Humanities and Computer Science (DHCS).

George H. John, Pat Langley. 1995. Estimating Continuous Distributions in Bayesian Classifiers. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, San Mateo, pages 338-345.

Thorste Joachims. 1999. Making Large Scale SVM Learning Practical. In *Advances in kernel methods*, page 184. MIT Press.

S. Sathiya Keerthi, Shirish K. Shevade, Chiranjib Bhattacharyya, and Karuturi R. K. Murthy. 2001. Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Computation*, 13 (3): 637-649.

Dan Klein and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, Volume 1: 423-430. Association for Computational Linguistics.

Kevin Knight. 1999. Mining Online Text, *Commun. ACM*, 42(11): 58-61.

Moshe Koppel, Shlomo Argamon, and Anat R. Shimoni. 2002. Automatically Categorizing Written Texts by Author Gender. *Lit Linguist Computing*, 17 (4): 401-412.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon 2011. Authorship Attribution in the Wild. *Language Resources & Evaluation*, 45(1): 83-94.

Mark L. Knapp, Roderick P. Hart, and Harry S. Dennis. 1974. An Exploration of Deception as a Communication Construct. *Communication Research*, 1: 15-29. doi:10.1111/j.1468-2958.1974.tb00250.x

Niels Landwehr, Mark Hall, and Eibe Frank. 2005. *Logistic Model Trees*. Machine Learning, 59 (1-2): 161-205.

Chul S. Lim, Kong J. Lee, and Gil C. Kim. 2005. Multiple Sets of Features for Automatic Genre Classification of Web Documents. *Information processing & management* (5): 1263-1276.

Dimitris Liparas, Yaakov HaCohen-Kerner, Stefanos Vrochidis, Anastasia Moumtzidou, and Ioannis Kompatsiaris. 2014. News Articles Classification Using Random Forests and Weighted Multimodal Features. In *Multidisciplinary Information Retrieval, Proceedings of the 7th Information Retrieval Facility Conference*, pages 63-75. Springer International Publishing.

Martins, B., Silva M. J. 2005. Language Identification in Web Pages. In *Proceedings of the 2005 ACM symposium on applied computing*, pages 764-768. ACM

Jacob Mey. 2001. Pragmatics: An introduction. Malden, MA: Blackwell Publishers.

Andrew McCallum and Kamal Nigam. 1998. A Comparison of Event Models for Naive Bayes Text Classification. In *Proceedings of the AAAI-98 workshop on learning for text categorization*, Vol. 752, pages 41-48.

Rada Mihalcea, and Carlo Strapparava. 2009. The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language. In *Proceedings of the ACL-IJCNLP 2009 Conference*, Short Papers, pages 309-312. Association for Computational Linguistics.

Myle Ott, Yejin Choi, Claire Cardie, Jeffrey T. Hancock. 2011. Finding Deceptive Opinion Spam by any Stretch of the Imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Volume 1, pages 309-319. Association for Computational Linguistics.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: Sentiment Classification using

Machine Learning Techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing* (EMNLP'02), Volume 10, pages 79-86.

Maria T. Pazienza. (ed.) 1997. Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology. Springer International Publishing.

James W. Pennebaker, Cindy K. Chung, Molly Ireland, Amy Gonzales, and Roger J. Booth. 2007. The Development and Psychometric Properties of LIWC2007.

John C. Platt. 1998. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. *Advances in Kernel MethodsSupport Vector Learning*, 208:1–21.

J. Ross Quinlan. 1993. Programs for Machine Learning. Volume 240. Elsevier.

J. Ross Quinlan. 2014. C4. 5: Programs for Machine Learning. Elsevier.

Fabrizio Sebastiani. 2002. Machine Learning in Automated Text Categorization. *ACM computing surveys (CSUR)*, 34 (1): 1-47.

Fabrizio Sebastiani. 2005. Text Categorization, pages 683-687. Retrieved from: http://nmis.isti.cnr.it/sebastiani/Publications/EDTA05.pdf

Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis. 2000. Automatic Text Categorization in Terms of Genre and Author. *Comput. Linguist*, 26 (4): 471-495.

Efstathios Stamatatos. 2009. A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for information Science & Technology (JASIST)*, 60 (3): 538-556.

Marc Sumner, Eibe Frank, and Mark Hall. 2005. Speeding up Logistic Model Tree Induction. In *Proceedings of the Knowledge Discovery in Databases*, PKDD 2005, pages 675-683, Springer Berlin Heidelberg.

Vladimir Vapnik. 2013. The Nature of Statistical Learning Theory. Springer Science & Business Media.

Ian H. Witten and Eibe Frank, E. 2005. Data Mining: Practical Machine Learning Tools & Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems). San Mateo, CA: Morgan Kaufmann.

Yiming Yang. 1999. An Evaluation of Statistical Approaches to Text Categorization. *Information Retrieval*, 1(1-2): 69-90.

Yiming Yang and Jan O. Pedersen. 1997. A Comparative Study on Feature Selection in Text Categorization. In *Proceedings of the 14th International Conference on Machine Learning*, 97: 412–420.