

How to Cite:

Gajbhiye, P., & Agrawal, A. J. (2022). Pragmatic approach for Twitter analysis and perform prediction. *International Journal of Health Sciences*, 6(S2), 9463–9476.
<https://doi.org/10.53730/ijhs.v6nS2.7479>

Pragmatic approach for Twitter analysis and perform prediction

Prajakta Gajbhiye

Computer Science & Engineering, Shri Ramdeobaba College of Engineering and Management, Nagpur, Maharashtra, India

Corresponding author email: gajbhiyepn@rknec.edu

Dr. Avinash J. Agrawal

Computer Science & Engineering, Shri Ramdeobaba College of Engineering and Management, Nagpur, Maharashtra, India

Email: agrawalaj@rknec.edu

Abstract---Study of Pragmatics is a very important part of Natural Language Processing (NLP). Pragmatic analysis allows to analyze what the given text basically means. The aim is to draw inferences from the given text. This paper reports the work on pragmatics analysis performed for the Twitter data set. To understand the intended meaning behind a tweet, sentiment analysis, polarity and contextual information is used. Various machine learning algorithms including Logistic Regression, Decision Tree, Random Forests, and Support Vector Machines are used for implementation. A comparative analysis of evaluation using this algorithm is also reported.

Keywords---adaptive language understanding, extracting information, natural language processing, pragmatic analysis, machine learning, sentimental analysis.

Introduction

Natural language processing (NLP) has emerged as an interdisciplinary area. It is a compound of computational linguistics and Artificial Intelligence (AI). NLP can make machines capable to understand, analyze and generate human languages. NLP applications includes: machine translation, automatic summarization, Named Entity Recognition (NER), speech recognition, relationship extraction, and topic segmentation. There are five basic phases of NLP as in Fig 1.



Fig 1: Five Phases of NLP

Lexical Analysis and Morphological: This scans the input code characters and converts them into meaningful lexemes. It distributes the full text into sentences, words, and paragraphs.

Syntactic Analysis (Parsing): This is often accustomed to check grammar in sentences, arrange words, and sort of relationship occurs among them.

Semantic Analysis: It represents the meaning representation. It focuses on the accurate meaning of sentences, phrases, and words.

Discourse Integration: Discourse includes the study of chunks of language that means nothing but a bigger sentence as compared to a single sentence. It also contains directly preceding sentences that mean to predict the next sentence. Discourse language is important for solving pronouns and temporal characteristics of the knowledge transferred.

Pragmatic Analysis: It helps you to get the particular meaning of sentences with the use of rules and also analysis the intention behind the text. For Example: "Do you know what time it is?" is interpreted as a scolding you, not asking about time.

Pragmatics is the study of the relation between language and the context of use. Context of use included such things as identities of the people and objects. Hence, pragmatics includes an analysis of how language is used to ask people and things. Pragmatic Analysis deals with the communicative and social content and its effect on interpretation. It means extracting or determining the meaningful use of language in situations. During this analysis, the most focus on what was said is reinterpreted on what's the particular meaning. It helps users to get this expected effect by applying a group of rules that define supportive conversations. Mapping is formed between the syntactic structures and objects within the task domain. This analysis deals with external world knowledge, which uses for documents and/or queries. This analysis focuses on what was described and is reinterpreted by what it meant, deriving the assorted aspects of language that need real-world knowledge. It's a part of the method of extracting information from text. For instance, treating the word "board" as a noun or verb? Major applications include Machine translation (MT), information extraction (IE), information retrieval (IR), sentiment analysis, and question and answering chat-box.

Related Work

Bhavesh Kumar et al. [1] propose a working model of Pragmatic Analysis over a sample set of sentences, chosen from British English, which could further be extended. This paper uses a machine learning approach of Neuro-Fuzzy interpretation is used to achieve an adaptive language understanding, with a focus on the Intentions of the speaker/writer and performing Pragmatics Analysis.

Brian W. Patterson et al. in 2019 [2] propose used of pragmatic NLP algorithm approach to identifying falls in older adults within the emergency department was ready to identify falls with excellent precision and recall, similar to that of more labor-intensive manual abstraction.

Michael Collins et al. [3] propose use Machine Learning Methods in language Processing like Information extraction Named entities, Relationships between entities, finding linguistic structure, Check Word functions in meaning still as grammatically within the sentence, and computational linguistics uses Perceptron Algorithm and solving Problem of Pragmatic Ambiguity.

Chris Cherpas et al. in 1992 [4] propose to use pragmatic analysis and verbal behavior for what an individual speaks whether it's negative or positive or neutral and supported that performs pragmatic analysis of language. The evolutionary nature of behavior suggests an AI technology referred to as genetic algorithms/programming for implementing such a system.

Xiaorong Luo et al. in 2011 [5] studied On Pragmatic Failures in Second Language Learning and the uses of language in communication, which has interpretation and use of utterances, depend upon knowledge of the world. Speaker uses and understands speech acts just like the conversation between the speaker and listener. If there any pragmatic ambiguity occurs in NLP then language is misunderstood and therefore the communication fails this can be the most reason for pragmatic failure. Pragmatic failures occur thanks to the various meanings of identical sentences

Shehdeh Fareh et al. in 2008 [6] studied corpus for extracting meaningful data from the gathering of texts and also attempts to spot the varied discourse functions of interrogative sentences. Corpus is one of the methods of NLP. Corpus uses a library like Natural Language Toolkit (NLTK) and performs many tasks for text analysis. Using the NLTK we can see how many number nouns, verbs, adjectives, and adverbs are present in the sentences.

Ayşe Pinar Saygin et al. in 2002 [7] study of pragmatics in human-computer communication. It should require some modifications of existing frameworks in pragmatics. While Pragmatics constitutes a serious challenge to linguistics.

Jihen Karoui. Et al.in 2017 [8] propose explains the application corpus for automatic irony detecting if a given tweet is ironic or not. The goal of the paper analyzes if these categories are Valid in social media content and to extract meaningful data from the tweet. Analysis Tweets how an individual reacts to

social media. Use a multi-layered annotation schema for irony in the tweeter set and a multi-lingual corpus-based study for measuring the impact of pragmatics.

Yan Li in 2020 [9] mainly focuses on Information Systems. For that use three NLP research perspectives first Natural Language Inference to Semantic annotation second Corpus Analysis to Semantic annotation and last Text Generation to Information retrieved.

Mirella Lapata in 2006 [10] proposes a data-intensive approach for inferring sentence-internal temporal relations. The temporal inference is relevant for practical NLP applications that either extract or synthesize temporal information (e.g, summarization, question answering).

Rickey E. Carter in 2009 [11] uses Artificial intelligence and deep learning methods gave high progress in medical science like in radiographic images. But the challenges, risk to patient privacy, reproducing results, questions regarding ownership, and financial value of large medical datasets.

Experimental Work

A system design is proposed to analyze pragmatics in twitter data. After reading the data a series of preprocessing functions needs to be applied. The purpose of preprocessing is to increase efficiency of the system. Fig [2] presents the work flow of the proposed system. Detailed explanation of each step is also included in this section.

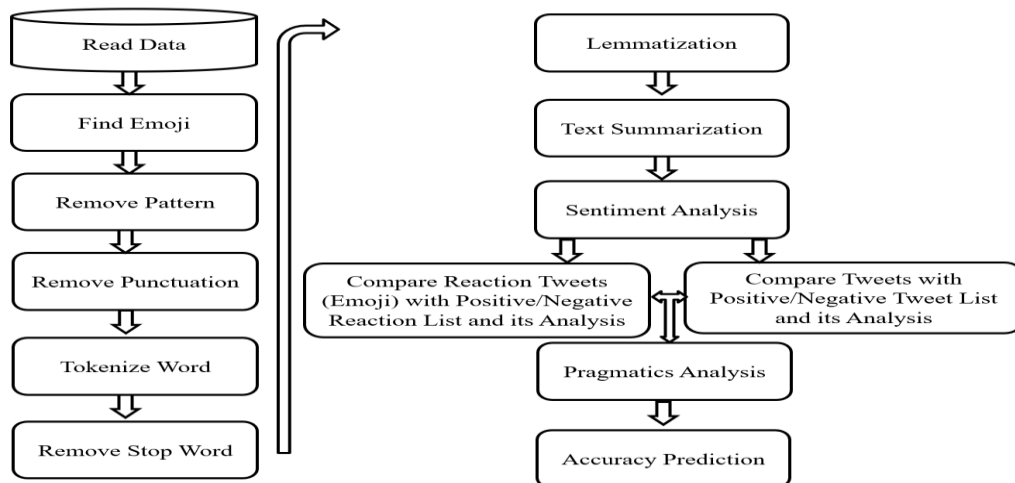


Fig 2:ProcessFlow

- A. Dataset: Get a Twitter data set in Excel by using Tweepy (Twitter API).Which is shown in fig [3].

- 2) Remove Punctuation: Remove tags from a given CSV file because they don't have value. Also, remove URLs (or Uniform Resource Locators) which are nothing but a location of the web pages and remove special characters like #(hashtag) and @(at the rate) this kind of symbol doesn't have any value, so we remove these from the CSV file. For Remove all kinds of special symbols we use the punctuation library in our programming.
 - 3) Tokenization sentence and word: Tokenization means separating a bit of text into smaller units called tokens. Here, tokens are often words, characters, or sub-words. Here we also Tokenize Alphanumeric from tweets. For performing tokenization use NLTK (Natural Language Toolkit) library.
 - 4) Remove stop words: Aside from URLs, HTML tags, and special characters, some words aren't required for tasks like sentiment analysis or text classification. Words like I, me, and you increase the scale of text data but don't improve results dramatically, and thus it's a decent idea to get rid of those. For the task, we will use a pre-defined stop words collection (e.g., from NLTK or the other NLP library), or we can define our own set of stop words supported by our task.
- D. Lemmatization: It's a way label for grouping many various sorts of words into the basic form which contains identical meanings. A word in an exceeding text may exist in multiple forms like stop and stopped (past participle) or price and costs (plural). This converts variations of the word into the basis variety of the identical word. For Lemmatization we can use WordNetLemmatizer from the NLTK WordNet library.
- E. Text Summarize: It means that obtain useful information from sources. While getting the information means fetching structural data from unstructured and/or semi-structured machine-readable documents. In this summary, we extract data using large texts to small texts. Here we use the gensim library to perform text summarization. Which reduces tweets or summarizes by 20%
- F. Sentiment Analysis: After getting Text Summarize, we perform Sentiment Analysis on it. It simply means recognizing the sense or sensation behind a situation whether a sentence was positive, negative, or neutral. The sentiment goes to calculate by using a lexicon. Lexicon theory means how the sentence is semantic oriented and how much intensity of every word within that sentence. For this, we need a dictionary that organizes between positive and negative words. Generally, a message is represented by several words. After assigning individual scores for all the words, the final sentiment is calculated by using three classifications polarity, subjective, and intensity.

These are the main 3 classifications required for calculation:

- 1) Polarity: negative vs. positive (Lies between -1.0 to +1.0)
- 2) Subjectivity: objective vs. subjective (Lies between 0.0 to +1.0)
- 3) Intensity: modifies next word (Lies between x0.5 to x2.0)

Polarity means some predefined restrictions or how much weight contain in a dictionary, whereas it has some limited score which lies between -1 to +1 and calculates a sentence's or tweet polarity. After getting the polarity score, we convert it into three parts whereas a negative score goes to the Negative tweet. A positive score or more than zero scores is called a positive tweet and if a score is zero then the tweet is neutral.

Subjectivity means sentences are based on a subject not depending upon non-public view and contain some factual information in the sentence itself. But the higher subjectivity provides personal opinions, not based on factual information. Intensity determines if a word modifies the following word. For English, adverbs are used as modifiers for example "very good".

For Sentiment Analysis, we use the Text Blob library. It gave the polarity and subjectivity of a sentence. Text Blob has a semantic label which gave us a very good analysis ex: emoticons, punctuation, emojis etc. Now we see step by step how the polarity and subjectivity calculate in sentences by using every word polarity and subjectivity. Lexicon refers to is in en-sentiment.xml. Here you can also see the Polarity, Subjectivity, and Intensity of words. If a word is repeated then check its sense and refer to its Polarity, Subjectivity, and Intensity. For Example: It is not a very good decision.

Step 1: Sentence "It is not a very good decision". Sentiment (polarity = -0.26, subjectivity = 0.46). This value gets by using Text Blob ("It's not a very good decision").sentiment

Step 2: word "decision". Sentiment (polarity = 0.00, subjectivity = 0.00). This value gets by using Text Blob ("decision") sentiment. The polarity and subjectivity are 0.00 because "decision" this word not present in the dictionary.

Step 3: word "good". Sentiment (polarity = 0.7, subjectivity = 0.60). This value gets by using Text Blob ("good") sentiment

Step 4: word "very". Sentiment (polarity = 0.2, subjectivity = 0.3, intensity = 1.3). This value gets by using Text Blob ("very") sentiment. Remember "very" is one of the modifier words then Text Blob just ignores the value of polarity and subjectivity and only focuses on the intensity of modifying the following word.

Step 5: word "very good". Sentiment (polarity = 0.90, subjectivity = 0.78). This value gets by using Text Blob ("very good") sentiment. For polarity it just Addition the value of very (polarity) and good (polarity) which is $0.2 + 0.7 = 0.90$ For subjectivity it uses the product between very (intensity) and good (subjectivity) which means $1.3 * 0.6 = 0.90$ Here "very" is a modifier word that specifies some meaning why we use the intensity of the word "very" with follow word "good" and calculate the subjectivity of sentences

Step 6: word "not". Sentiment (polarity = 0.00, subjectivity = 0.00). This value gets by using Text Blob ("not") sentiment. The polarity and subjectivity are 0.00 because the "not" this word is not present in the dictionary.

Step 7: word “not a very good”. Sentiment (polarity = -0.26, subjectivity = 0.46). Now “not” is a change in the whole meaning of sentences that is why we use -0.5 for a polarity. For calculating sentence polarity take a product of (-0.5) for “not”, the inverse of the intensity word “very” and the polarity of the word “good” which is $-0.5 * (1/1.3) * 0.7 = -0.26$. For calculated subjectivity take the inverse of the intensity word "very" and the product with the subjectivity of the word "good" which is $(1/1.3) * 0.6 = 0.46$

By using this analysis, we get a pie chart which is shown in Fig.5. For performing Sentiment Analysis in tweeter data set. Here we use Text Blob Library.



Fig 5: Sentimental Analysis in Twitter Dataset

G. Pragmatics Analysis: After getting a sentimental analysis we perform pragmatics on it. Pragmatics means the intention behind the text and what users' tweet. We also called sarcasm tweets or ironic tweets. A person uses those words that represent the opposite of what the person wants to text or speak, mainly to humiliate someone, display anger, irritate someone, or simply to be funny. For performing pragmatics, we create a simple logic.

- 1) Compare “Reaction Tweets” (Emoji) with “Positive Reaction List” and its “Analysis”: Here we create a “positive reactions list” and check whether the “Tweet reaction” contains that list or not. If the “reaction” contains then we also check whether the “Analysis” is negative or not. If it's negative then those “tweets” come under Pragmatics. If a person posts a negative “tweet” then highly chance it posts an Angry Face emoji, not a joyful face Emoji we predict that “tweet” is pragmatic.

Reaction” if the matching is present then we also match its tweets “Analysis” if it's Positive then the “tweet” is Pragmatics.

- 3) Compare “Tweets” with “Positive Tweet words List” and its “Analysis”: Now we create a “positive tweet words list” and check whether the “Tweet” contains those words or not. If the “Tweet” contains those words then we also check whether the “Analysis” is negative or not. If it's negative then those “tweets” come under Pragmatics. If a person posts a negative “tweet” then highly chance it contains some sad words and de-motivational words, not happy words, or encouraging words, so we predict that “tweet” is pragmatics.

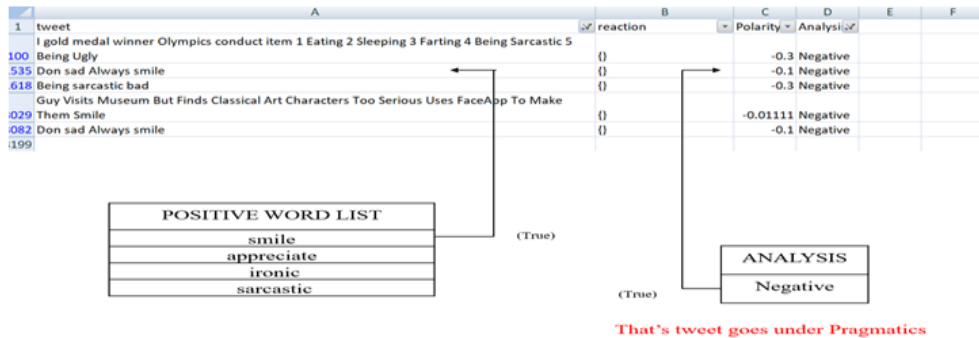


Fig8:Match Positive Tweet word List with Tweets and Tweet Analysis

In this Fig [8] first, we create a “Positive word List” which contains four words “smile”, “appreciate”, “ironic” and “sarcastic”. After then we match with “tweets” and if the matching is present then we again match its tweet “analysis”. If it’s negative then the tweet is Pragmatics.

- 4) Compare “Tweets” with “Negative Tweet words List” and its Analysis: Now we create a “negative tweet words list” and check whether the “Tweet” contains those words or not. If the Tweet contains those words then we also check whether the “Analysis” is positive or not. If it's positive then those “tweets” come under Pragmatics. If a person posts a positive “tweet” then highly chance it contains some motivational words and best wishes words, not sadwords, so we predict that “tweet” is pragmatics.

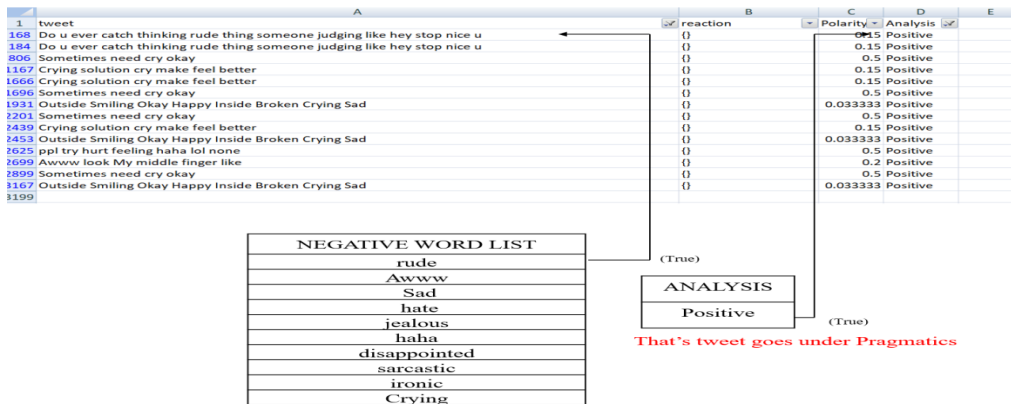


Fig 9: Match Negative Tweet word List with Tweets and Tweet Analysis

In this Fig [9] first, we create a “Negative word List” which contains ten words like “rude”, “Awww”, “Sad”, “hate”, “jealous”, “haha”, “disappointed”, “sarcastic”, “ironic”, “Crying”. After then we match with “tweets” and if the matching is present then we again match its tweet “analysis”. If it’s Positive then the tweet is Pragmatics.

ANALYSIS	POSITIVE REACTION LIST	NEGATIVE REACTION LIST
Negative	Pragmatics	Non Pragmatics
Positive	Non Pragmatics	Pragmatics

ANALYSIS	POSITIVE WORD LIST	NEGATIVE WORD LIST
Negative	Pragmatics	Non Pragmatics
Positive	Non Pragmatics	Pragmatics

Table 1: Logic find tweet pragmatics or not

In this Table (1) we can see the logic behind message pragmatics or not. If “Positive Reaction List” and its “Analysis” are negative then it comes under Pragmatics. If “Negative Reaction List” and its “Analysis” are positive then it comes under Pragmatics. Same as if “Positive word list” and its “Analysis” are negative then it comes under Pragmatics. If “Negative Word List” and its “Analysis” are positive then it comes under Pragmatics.

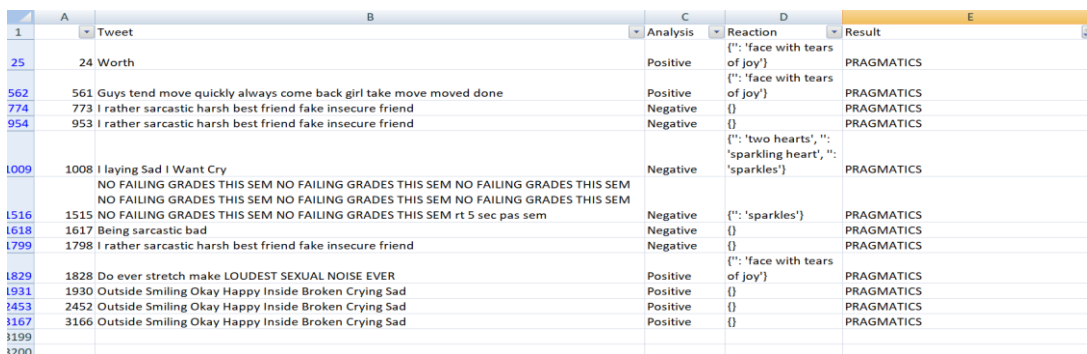


Fig10: Tweet Pragmatics or not

In this Fig (10) using this logic, we get whether the tweet is pragmatic or not. These are some tweets that contain pragmatics.

H. Accuracy Prediction: In this Accuracy prediction simple we use four Algorithms to find accuracy between Tweets and results. Here are those four Algorithms.

- 1) Logistic Regression: Logistic regression is a considerable Machine learning Technique that uses Supervised Learning. It operated a predicting unconditional dependent value by using two independent sets. Logistic regression predicts the accuracy uses of dependent values.
- 2) Decision Tree: In that Decision tree, it is also a Supervised Machine Learning where data is split between nodes and leaves.
- 3) Random Forests: Random Forest is one of the Supervised Machine Learning techniques which is use Classification and Regression problems. It makes a Decision tree on another sample set and takes the majority value for classification and regression using the average case.
- 4) Support Vector Machine (SVM): SVM also uses a supervised technique. This algorithm uses both regression and classification.

Algorithm	Test Accuracy	Train Accuracy	Mean Accuracy	Std of Accuracy	Auc for our data
Logistic Regression	0.9990	0.9951	0.9951	0.0013	(0.5, 3)
Decision Tree	0.9979	1.0000	0.9969	0.0020	(0.0,0)
Random Forests	0.9990	1.0000	0.9978	0.0022	(0.5, 3)
Support Vector Machines	0.9990	0.9982	0.9978	0.0022	(0.5, 3)

Fig 11. Accuracy Prediction

In this Fig (11) we can see that our analysis according to Algorithms. Our best Accuracy is 0.9990 for Testing data set which is given by Logistic Regression, Random Forests and Support Vector Machines.

Conclusion

Pragmatics is a study or analysis field with many ideas for research. Uses of Pragmatics analysis what is meaning of the message or text can be predicted. The uses of human different languages encourage the creation of new Algorithms which make machines more intelligent for understanding humans' languages. Pragmatics sometimes deals with the impressions of text or message. As more and more algorithms for capturing context are made, the accuracy of algorithms implementing pragmatics. Pragmatic information is valued as one of the most difficult languages and comes only through experience.

References

1. Bhavesh Kumar, Hima Bindu Maringanti, Krishna Asawa , "Adaptive Pragmatic Analysis of Natural Language" Jaypee Institute of Information Technology A-10, Sector-62, Noida.
2. Brian W. Patterson , Gwen C. Jacobsohn¹, Manish N. Shah , Yiqiang Song, Apoorva Maru, Arjun K. Venkatesh, Monica Zhong, Katherine Taylor, Azita G. Hamedani¹ and Eneida Mendonc² "Development and validation of a pragmatic natural language processing approach to identifying falls in older adults in the emergency department.
3. Michael Collins "Machine Learning Methods in Natural Language Processing" MIT CSAIL Columbia University Department of Computer Science.
4. Chris Cherpas "Natural language processing, pragmatics, and verbal behaviour" by sharps Fremont, CA. 1992; 10:135-47. doi: 10.1007/BF03392880.
5. Xiaorong Luo "On Pragmatic Failures in Second Language Learning" ISSN 1799-2591 Theory and Practice in Language Studies, Vol. 1, No. 3, pp. 283-286, March 2011.
6. Shehdeh Fareh and Maher Bin Moussa "Pragmatic Functions of Interrogative Sentences in English: A Corpus-based Study" University of Sharjah International Journal of Arabic-English Studies (JJAES).
7. Ayse Pinar Saygin And Ilyas Cicekli "Pragmatics In Human-Computer Conversations" Of cognitive Science, Univ. Of California, San Diego, La Jolla, CA 92093-0515, USA Dept. of Computer Engineering, Bilkent University, 06533 Bilkent, Ankara, Turkey.
8. Jihen Karoui, Farah Benamara, Veronique Moriceau, Viviana Patti, Cristina Bosco, and Nathalie Aussenac-Gilles "Exploring the Impact of Pragmatic Phenomena on Irony Detection in Tweets: A Multilingual Corpus Study" University of Turin, Italy.
9. Yan Li , Manoj A Thomas & Dapeng Liu "From Semantics To Pragmatics: Where IS Can Lead In Natural Language Processing (NLP) Research" European Journal Of Information Systems.
10. Mirella Lapata, Alex Lascarides "Learning Sentence-internal Temporal Relations" Journal of Artificial Intelligence Research 27 (2006) 85–117.
11. Rickey E. Carter , Zachi I. Attia, Francisco Lopez-Jimenez and Paul A. Friedman "Pragmatic considerations for fostering reproducible research in artificial intelligence" nature parter journal
12. "Pragmatics and natural language generation" by Eduard h. hovy information sciences institute of the University of Southern California 4676 admiralty way marina Del Rey, ca 90292-6695 U.S.A.
13. "An example of pragmatic analysis in natural language processing:" sentimental analysis of movie reviews communication and technology congress – © Istanbul Ayden university.
14. "Formal semantics and pragmatics for natural language querying" James Clifford (New York university) Cambridge, England: Cambridge University. (Cambridge tracts in theoretical computer science 8) hardbound, ISBN.
15. "Fundamentals of natural computing: an overview" Leandro Nunes de Castro graduate program in computer science, catholic university of Santos, r. dr. carvalho de mendonça, 144, vila mathias, santos sp, brazil communicated by I. Perlovsky.

16. "Semantics and pragmatics and future challenges in NLP" by mark Granroth-wilding.
17. "Pragmatic approach in natural language understanding" by Shinji Mitsubishi, fuji ren, yalin, john r. Ogawa.