Indian Institute of Technology Bombay

Research & Development Project

Cross Lingual Sarcasm Detection

Author:
Siddharth Bulia

Supervisor:
Dr. Pushpak Bhattacharyya

A thesis submitted in fulfillment of the requirements for Research and Development Project

Department of Computer Science and Engineering
Indian Institute of Technology Bombay

April 30, 2016

Acknowledgements

I would like to take this opportunity to express my deepest gratitude towards my guide *Dr. Pushpak Bhattacharyya* for his motivating support, guidance and valuable suggestions. I would also like to thank *Aditya Joshi* from the *Sentiment Analysis Team* at IIT Bombay who helped me throughout the project and provided constructive feedback.

Abstract

Sentiment Analysis Research has reached to a altogether new level with the advent of web 2.0. Sentiment Analysis has found its utility in a number of sectors. Industry is making use of sentiment analysis in various fields like analysing the reviews, personal assistance, medication, etc.
With the increase in use of sentiment analysis, there is an urgent to handle the problems arising in this field. One of the major problem is sarcasm detection. While there are ways to deal with it in highly used languages like English, still there is very less amount of research done in detecting sarcasm in less popular languages like Italian, Czech, etc. because of unavailability of resources.
In this report, I present a brief survey of the fascinating research going on in sentiment analyis. I also discuss my work in collaboration with Aditya Joshi on sarcasm detection in less popular languages where resources are scarce using Cross Lingual Sarcasm Detection.

Contents

Chapter 1

Introduction

The rise of Web 2.0 has resulted in a lot of increase in online digital data. Web 2.0 site allow users to interact and collaborate with each other in a social media dialogue as creators of user-generated content in a virtual community, in contrast to Web sites where people are limited to the passive viewing of content. People are participating in creating content over the web through many forms like blogs, reviews, social media, etc. This content contains opinion of the writers about the various issues going over in the country, about products, people, etc. The value of opinion on the web may be looked at from two perspectives: business and social. The business of an organization may benefit by tapping opinion expressed by internet users about their products/services. Additionally, businesses may wish to know, in real-time, how their decisions and public announcements have been received by people on the web. The social perspective of online opinion can be understood from the fact that it has mobilized real-world events in the past. Specifically, in recent times, two public movements driven by social media took place in India. These events, namely an anti-corruption demonstration 1 and protests following a sexual assault incident 2 , highlight the impact of web opinion on real world events in today's times. The rising impact of digital opinion has accelerated research in sentiment analysis.

Sentiment Analysis researchers face a number of challenges, one of which is Sarcasm Detection. Given a sarcastic sentence, if a normal sentiment analysis engine try to predict its sentiment, the results come out as opposite of what is expected. Researchers are working in the field of sarcasm detection to overcome this challenge. But the languages which are less popular in terms of research, example hindi, marathi, where we don't have enough resource. Normal Sarcasm Detection/Sentiment Analysis methods does not work here. In such cases we use Cross Lingual Sentiment Analysis/Cross Lingual Sarcasm Detection is used.

This project is done in collaboration with Aditya Joshi, member of sentiment Analysis Team at IIT Bombay. The upcoming chapter introduces sentiment analysis, cross-lingual sentiment analysis, sarcasm detection, cross lingual sarcasm detection and our experiments done in Cross Lingual Sarcasm Detection. Chapter 2 defines sentiment analysis (SA) and describe its challenges, approaches and lexical resources. Chapter 3 describe the cross lingual sentiment analysis and present the challenges, approaches and corpora required for cross-lingual SA. Chapter 3 defines Sarcasm Detection and its benefits and in chapter 5, we introduce a new field Cross lingual Sarcasm detection and our experiments in this field. Last Chapter, chapter 6 deals with the observations arising from experiments, error analysis and future work.

Chapter 2

Sentiment Analysis

According to Bing Liu [1], *sentiment analysis*, also called *opinion mining*, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. It represents a large problem space. There are also many names and slightly different tasks, *e.g., sentiment analysis, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, etc.* However, they are now all under the umbrella of sentiment analysis or opinion mining. While in industry, the term sentiment analysis is more commonly used, but in academia both sentiment analysis and opinion mining are frequently employed.

There are more than one definitions defined for sentiment analysis. Apart from research papers, several websites give their own definition in simple words.*Sentiment analysis* is the process of determining whether a piece of text is positive, negative or neutral.[].

Sentiment analysis refers to the use of natural language processing, text analysis, computational linguistics to identify and extract subjective information in source material.[]

For example, The term sentiment analysis was first used by Nasukawa and Yi, 2003 [2].However, the research on sentiments and opinions had begun earlier (2000). Although linguistics and natural language processing (NLP) have a long history, little research had been done about people's opinions and sentiments before the year 2000. Since then, the field has become a very active research area. There are several reasons for this:

- It has a wide arrange of applications, almost in every domain. The industry surrounding sentiment analysis has also flourished due to the proliferation of commercial applications. This provides a strong motivation for research.

- It offers many challenging research problems, which had never been studied before.
- For the first time in human history, we now have a huge volume of opinionated data in the social media on the web. Without this data, a lot of research would not have been possible. Not surprisingly, the inception and the rapid growth of sentiment analysis coincide with those of the social media. In fact, sentiment analysis is now right at the center of the social media research.
- Research in sentiment analysis not only has an important impact on NLP, but may also have a profound impact on management sciences, political science, economics, and social sciences as they are all affected by people's opinions.

2.1 Challenges

Not surprisingly, the most important indicators of sentiments are *sentiment words*, also called opinion words. These are words that are commonly used to express positive or negative sentiments. For example, *good, wonderful* and *amazing* are positive sentiment words, and *bad, poor* and *terrible* are negative sentiment words. Apart from individual words, there are also phrases and idioms, e.g., *cost someone an arm and a leg.* Sentiment words and phrases are instrumental to sentiment analysis. A list of such words and phrases is called a *sentiment lexicon (or opinion lexicon).*

Although sentiment words and phrases are important for sentiment analysis, only using them is far from sufficient. The problem is much more complex. In other words, we can say that sentiment lexicon is necessary but not sufficient for sentiment analysis. Below, we highlight several issues:

- **Domain Dependent:** Sentiment Analysis highly depend upon the domain in which the sentence is used. A sentence can be considered as positive in one domain but might be considered as negative in another domain.

  Example: *Go read the book.*

  If we come across this sentence in a book domain, we can infer that the writer is giving positive review about the book. He/She wants the reader to read the book. While if we come across this sentence in a movie domain, the author might want to tell the reader to read the book instead of watching the movie. Hence it is negative review.

- **Sarcasm:** *Activity of saying or writing the opposite of what you mean, or of speaking in a way intended to make someone else feel stupid or show them that you are angry* (Macmillan English Dictionary 2007).

As the actual meaning expressed by the author is opposite of what the sentence means, it is very harder for normal sentiment analysis engines to predict the actual sentiment of the author.

Example: *What a great car! It stopped working in two days*

The above sentence is a sarcastic one. If given to sentiment classifier directly it may predict it to be positive because of word "great" but it is not what the speaker meant.

- **Thwarted expression:** *Expression which contain a number of words that have polarity which is opposite to the polarity of the expression itself.*

  Example: *Johnny depp was alright. The previous two pirate were unrealistic and boring. The plot was awful. However the special effects made the third pirate movie excellent.*

  Here the sentence contain a large number of negative words but the overall opinion expressed in the review is positive. Such expressions can be easily found in reviews where there is comparison between a good product and bad.

- **Negation:** *Expression composed of negating word (like not, neither) followed by a noun, adjective, adverb or verb.*

  Example: *The movie was not enjoyable neither entertaining.*

  The overall polarity of the expression is negative but it is very hard for a sentiment analyzer to guess because of presence of positive words.

- **Implicit Polarity:** *Expressions not containing polar words but have a overall opinion.*

  Example: *The handle breaks too easily.*

  The above statement seem to contain some sentiment regarding the handle but because it does not contain any polar word, it is difficult to predict its polarity automatically.

2.2 Approaches in Computation Sentiment Analysis

- **Lexicon-based approaches:** A lexicon containing sentiment information of words is used to determine sentiment in text. These approaches may use a rule-based classifier to predict the output label.
- **Corpus-based approaches:** A classifier may be trained on corpus annotated with output labels. Supervised approaches used for SA differ in aspects including but not limited to feature representation, classifier used, etc.

2.3 Lexicons

Sentiment lexicons are resources that contain linguistic units annotated with sentiment information. These linguistic units may be words, phrases or word senses. In addition, sentiment information could be basic output labels, subjectivity information. Some commonly used lexicons for SA are:

- **SentiWordnet:** SentiWordnet given by Esuli and Sebastiani [2006] is a lexical resource that adds sentiment information to Wordnet in the form of three scores: positive, negative and neutral. The scores in Sentiwordnet have been assigned as a result of an automatic sentiment classifier trained on seed synsets. The sentiment annotation at synset level allows distinguishing between different senses of words that may have different polarity.
- **WordNet Affect:** Wordnet Affect is a lexical resource that allows determination of affective content of synsets by dividing them into affective categories. Thus, it gives more information as compared to SentiWordNet and is used when analysis to be done is with respect to emotions like anger, joy, etc.
- **Word Lists:** Word-lists like Opinion Finder by Riloff et al. [2005] are also popular. These lists differ in the input (word/sense/phrase, etc.) as well as in terms of the sentiment information. The annotation could be one or more out of 'positive/negative', 'positive/negative/neutral', 'strongly positive/weakly positive', etc.

Chapter 3

Cross Lingual Sentiment Analysis

Sentiment Analysis in a new language can done two ways. Developing lexicons, algorithms, classifiers for new language or by making use of lexicons, algorithms and classifiers developed in another language. The second method is known as Cross Lingual Sentiment Analysis.

Cross-Lingual Sentiment Analysis (CLSA) is the task of predicting the polarity of the opinion expressed in a text in a languae L_{test} using a classifier trained on the corpus of another language L_{train}. [3]. L_{test} is called the Target Language and L_{train} is called the Source Language.

Example: Sentiment Analysis in Sanskrit using the lexicons developed in Hindi. Sarcasm Detection in Italian using the annotated documents in English.

3.1 Challenges

- **Less Reliable Machine Translation:** Primary tool for doing Cross Lingual Sentiment Analysis is using a machine translation tool which converts the Source Language Documents into a Target language. A lot of research is happening in the field of machine translation but still machine translation is not very much reliable.

  English to hindi: *Today is awesome.* - आज भयानक है ।

  In the above sentences, translation completely changed the sentiment expressed in the original sentence. The second statement before translation expresses a positive sentiment but after translation it expresses a negative sentiment.

- **No Polarity:** There exist many words in a language which contain some polarity within them but on their translation, they lose it.

  English to Romanian: *fragile - fragil*

  In english, *fragile* carry some sentiment of softness while no such sentiment exist in Romanian word *fragil*

  Hindi to English: भैया *-Brother*

  In hindi, चाचू *chachu* carry some sentiment of love,respect while no such sentiment exist in English word *uncle.*

  In hindi, *chachu* carry some sentiment of love,respect while no such sentiment exist in English word *uncle.*

- **Limited Vocabulary:** Assume the case of English-Hindi cross-lingual SA. If a classifier is trained on documents translated from Hindi to English, it is possible that these documents contain words that are not commonly used in natural English sentences.

  Hindi: जूठा

  In hindi, जूठा means half-eaten food by a person. It does not have a corresponding word in english.

- **Instance mismatch:** It happens many a times that a word has more than one synonyms. When using a machine translation system, it is possible that for training, it converts all synonym words of one language into a particular word of second language. Now, it is possible that the test document does not contain that word but contain synonyms of it. In such cases, the classifier will ignore these words as they don't occur in its feature space.

  English to Hindi: *translation rocks* - अनुवाद चट्टान!

  In the former case, words *rocks* is used to tell how good translation is while in later, there meaning is completely changed.

- **Idioms:** Idioms or multi word phrases present in one language may not be present in other language. Or even if idiom exists, one should be able to directly map them from one language to another otherwise it might lose its polarity on translation.

  Consider idiom in English *foot in mouth*, it contains negative sentiment. Its proper translation in hindi is गलत कहने के बाद पछतावा करना but if we translate idiom word by word, we will get पैर में मुंह. The later sentence does not carry the same opinion as expressed in the english meaning.

3.2 Approaches used in Cross Lingual Sentiment Analysis

Basic technique used in cross lingual sentiment analysis is cross-lingual projection. In simple words, it means projecting words from vocabulary space to other vocabulary using some tools like Machine Translation, Bilingual Mapping, etc. Because there are some error in projection techniques, it causes some error in analysis as well.

Mainly there are three approaches used in CLSA :

- **Learning the classifier in source language space**

  In this method, the classifier is learnt over the training corpus in their original source language. To test any document of target language, the test document is first translated to source language and then tested. Its benefits are that the classifier is learnt over the natural source language sentences than on the translated sentences which may not be as fluent as well as adequate. While, one of the key negative point is while testing any document, even after training, you need to convert it into source language which is an extra overhead.

- **Learning the classifier in target language space**

  In this method, training data is first converted to the target language and then classifier is learnt over it. This classifier is used later on. Its key benefit is that one does need any translation method once the training is complete.

- **Learning the classifier in common language space**

  This category include those methods in which classifier is not learnt over on source language neither target language. Both the training and test corpus are first projected to some other common feature space like wordnet or third language, etc. And then later classifier is learnt over that common feature space. It can be used for those languages where direct conversion to each other does not exist but can be translated to a third commonly used language like english.

3.3 Translation for Cross Lingual Sentiment Analysis

Mainly the following techniques are used for translation in CLSA :

- Bilingual Dictionaries and Parallel Corpora

Using a well trained machine translation tool to translate data. Need to be done per sentence or per document. Examples . Google translate, Bing translate.

- Machine Translation

  Using a well trained machine translation tool to translate data. Need to be done per sentence or per document. Examples . Google translate, Bing translate.

Chapter 4

Sarcasm

Sarcasm is defined as *'the activity of saying or writing the opposite of what you mean, or of speaking in a way intended to make someone else feel stupid or show them that you are angry'*(Macmillan 2007) Another definition goes by *'a cutting, often ironic remark intended to express contempt or ridicule"*(free dictionary)

Sarcasm is present all over the internet, ranging from product reviews to social media. In normal day to day life as well, people use sarcasm all the time. By definition of sarcasm, the intended meaning of sarcastic sentences is opposite of what is written in the sentence. Hence sarcastic sentence has a large potential to defeat the sentiment analysis systems. To prevent this, a large amount of research is being conducted in the field of sarcasm detection. This chapter introduce sarcasm, examples of sarcasm in different contexts, need of sarcasm detection, work done in sarcasm detection, **Cross Lingual Sarcasm Detction**.

4.1 Examples of Sarcsasm

- *[I] love the cover* (book)
- *Wow, this thing makes me feel like I am in USA :p* (GPS device)
- *Be sure to save your purchase receipt* (smart phone)
- *Are these iPods designed to die after two years?* (music player)

The first sentence does not appear to be sarcastic if one does not look at the context. It appears like a genuine compliment, but in book domain there is a saying "Never judge a book by its book". Hence refering to this context, sentence appear sarcastic.

The second sentence also appears to a genuine compliment if one does not what is that "thing", but if one look at the context, one can easily understand the person is making sarcastic comment on a GPS device which is not functioning properly. ":p" smiley supports its sarcastic nature.

Third sentence come at the borderline of sarcasm and genuine advice.

In fourth sentence the sarcasm emerges from the naive-like question phrasing that assumes the general expectation that goods should last.

Hence, by observing these sarcastic sentences, one can observe that sarcasm comes in its own flavours. Sarcasm detection methods must have different featuresas compared to normal sentiment analysis method.

4.2 Benefits of Sarcasm Detection

- **Benefits of Sarcasm Detection:** Studies of user preferences suggest that some users find sarcastic reviews biased and less helpful while others prefer reading sarcastic reviews (the 'brilliant-but-cruel' hypothesis (Danescu-Niculescu-Mizil et al. 2009)). Identification of sarcastic reviews can therefore improve the personalization of content ranking and recommendation systems such as (Tsur and Rappoport 2009).

- **Improvement of Review Summarization and opinion mining :** Current opinion mining systems are incapable of dealing with sarcastic sentences. Sarcasm, at its core, may harm opinion mining systems since its explicit meaning is different or opposite from the real intended meaning (see example 1), thus averaging on the sentiment would not be accurate.

Chapter 5

Cross Lingual Sarcasm Detection

Cross-lingual sentiment analysis has been reported to be worthwhile for resource-scarce languages. The goal of this project is to validate if this holds in case of a related task: **sarcasm detection**

Cross Lingual Sarcasm Detection is an untouched area in which no research has been done so far. I, with Aditya Joshi from Sentiment Analysis Group, IIT Bombay, have attempted to do some research in this area.

5.1 Need of Cross Lingual Sarcasm Detection

- Cross Lingual Sarcasm Detection is as important as Cross Lingual Sentiment Analysis.
- With the increasing use of sentiment analysis in non-popular languages, problem arising in them need to be addressed one of which Sarcasm Detection.
- From analysis point of few, finding which languages are primarily used for sarcasm is a interesting topic.
- Developing complex NLP tools in other languages is a tough task.
- Majority of research done in Sarcasm Detection is in English language. To use that pre-existing research done in this field, cross linguality is required.

5.2 Challenges faces in Cross Lingual Sarcasm Detection

- Cross Lingual Sarcasm Detection bring up the problems of both areas, Cross Lingual Sentiment Analysis and Sarcasm Detection.

- Translation methods used for translation may not be reliable.
- Finding sarcasm in natural language is itself a tough job. Translation can completely change the structure/meaning of sentence.
- Idioms used in Sarcasm. Idioms used for sarcasm in language may not be used in other languages.
- Words used to show sarcasm in one language might not present in other one.
- Examples of challenges faces during translation :
  - *Yeah, today's morning is as awesome as that of 9/11.* - हाँ , आज की सुबह 9/11 की है कि जैसे ही भयानक है।
  - *I love being ignored.* - मैं अनदेखा किया जा रहा प्यार करता हूँ।
  - ***Person A*** - *Oh! Power is back.*
    Person B *: Oh? Really? But the tube-light still took time to light-up!*
    Person A - ओह! बिजली आ गई है।.
    Person B *:* ओह? वास्तव में? लेकिन ट्यूब प्रकाश अभी भी प्रकाश –अप करने के लिए समय ले लिया !

5.3 Development Work

Implemented a experimental setup and Web based User Interface.

- Implemented In Language and Cross Lingual Sarcasm Detector various NLP and ML tools.
- Developed a Web application which can be used directly from the browser to use the classifier learnt using the above setup. Web Application allow the user to give input in various languages including English, Italian, French, Hindi, etc.

5.4 Experimental Setup

5.4.1 Classifier

- **SVM** or **Support vector machine** with RBF kernel is used thoughout the project.

- After a lot of experimentation, the value of parameters, gamma and C is set to 0.3 and 1. **gamma** the gamma parameter defines how far the influence of a single training example reaches, with low values meaning 'far' and high values meaning 'close'. The gamma parameters can be seen as the inverse of the radius of influence of samples selected by the model as support vectors.The **C** parameter trades off misclassification of training examples against simplicity of the decision surface. A low C makes the decision surface smooth, while a high C aims at classifying all training examples correctly by giving the model freedom to select more samples as support vectors.

5.4.2 Pre Processing of Data

- Replaced all kinds of emoticons with their corresponding meanings. Emoticons with same meaning are clubbed together.
- Words carrying same type of meaning are replaced with their common meaning. Example: yay,yaay,yaay are replaced with good.
- Improper words used in tweets in tweets are replaced by their corresponding full words. Example: 'r' goes to 'are',"don't" goes to "do not",etc.
- Each exclamation and question mark symbol replaced with " Exclamation " and " QuestionMark ". If their are 3 (!!!) exclamation makrs present in a sentence, the corresponding processed tweets will contain 3 Exclamation word.
- Stemmers of Italian, English and Czech are used. English, Italian stemmers are taken from nltk library; Czech stemmer is taken from open source code by Luís Gomes. Stemming is the term used in linguistic morphology and information retrieval to describe the process for reducing inflected (or sometimes derived) words to their word stem, base or root form.

5.4.3 Features used

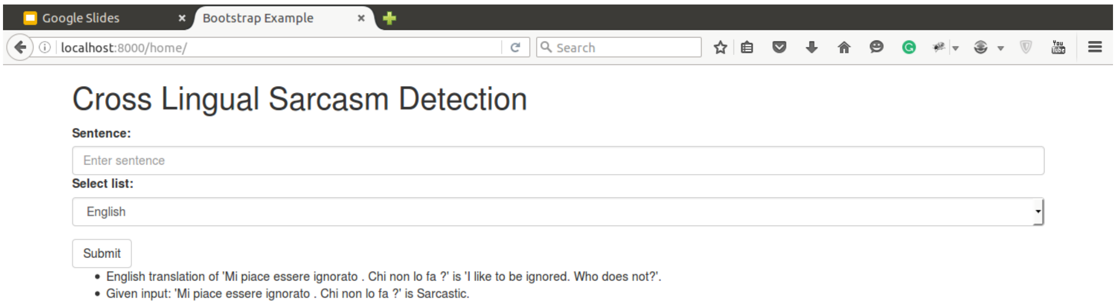
- 1-gram, 2-gram
- Emoticons
- Contradiction in statement
- Number of question mark
- Number of Exclamation mark

- Subjectivity of the tweet
- Length of the tweet

5.4.4 Translation methods Used

- Cross Lingual Sarcasm detection using Bilingual Mapping
- Cross Lingual Sarcasm Detection using Machine Translation

5.4.5 Web Application Interface



5.4.6 Approaches Used

- Learning the classifier in Source Language.
- Learning the classifier in Target Language.

5.5 Results

Each Experiment can be characterised by a 4-tuple - (S,T,A,M) where

- S means the Source Language. S can be English, Czech or Italian in this experiment.

- T means the Target Language. T can be Czech or Italian in this experiment. T as English is not required as English is not a resource scarce Language.
- A means the approach Used. A can Learning the classifier in Source Language (C_S) or Target Language(C_T).
- M means the Translation Method Used. M can Bilingual Mapping(BM) or Machine Translation(MT).

5.5.1 Experiment 1: Target Language Italian

Source: English, Czech
Approach: Classifier learnt in English, Italian or Czech

In Language		
Precision	Recall	F-Score
0.80	0.79	0.79

Source	Approach	Bilingual Mapping			Machine Translation		
		Precision	Recall	F-Score	Precision	Recall	F-Score
English	Italian	0.56	0.68	0.62	0.51	0.82	0.63
English	English	0.53	0.89	0.67	0.51	0.91	0.65
Czech	Italian	0.44	0.66	0.53	0.44	0.67	0.53
Czech	Czech	0.46	0.74	0.57	0.48	0.62	0.62

5.5.2 Experiment 2: Target Language Czech

Source : English , Italian
Approach : Classifier learnt in English, Italian or Czech

In Language		
Precision	Recall	F-Score
0.63	0.60	0.62

Source	Approach	Bilingual Mapping			Machine Translation		
		Precision	Recall	F-Score	Precision	Recall	F-Score
English	Czech	0.47	0.82	0.60	0.47	0.83	0.60
English	English	0.47	0.90	0.62	0.47	0.88	0.61
Italian	Czech	0.49	0.91	0.63	0.48	0.87	0.62
Italian	Italian	0.49	0.94	0.64	0.48	0.95	0.64

Chapter 6

Observation, Con clusions and Future Work

In the previous chapter, we presented our attempt to use Cross Linguality in area of Sarcasm Detection. We can extract out a few observations out of it.

6.1 Observations

- The results obtained in Cross Lingual Sarcam Detection are lower as compared to the results obtained through In Language Sarcam Detection.
- Results obtained using Bilingual Mapping and Machine Translation are almost equal with Machine Translation on a slightly better side.
- Results obtained in Italian are better than the results obtained in Czech.
- On training Classifier on source language gives better results.
- CLSD is not able to perform as good as CLSA.

6.2 Error Analysis

Target Language: Italian, Source Language: English, Approach: Classifier trained in Source Language

- Consider this Italian sentence : *giugn arriv commodor 64. dann poste.* #sarcastic

- Converted to English using Bilingual Mapping : *TO June arrives the Commodore 64. damned posed.* Classified as Non Sarcastic.

  Significant words present : commodore, arrives, damned, posed.

  Frequency of commodor* words in English Sarcastic sentence : 0

  Frequency of commodor* words in English Non Sarcastic sentence : 0

  Frequency of arriv* words in English Sarcastic sentence : 2

  Frequency of arriv* words in English Non Sarcastic sentence : 5

  Frequency of damn* words in English Sarcastic sentence : 24

  Frequency of damn* words in English Non Sarcastic sentence : 30

  Frequency of pose* words in English Sarcastic sentence : 0

  Frequency of pose* words in English Non Sarcastic sentence : 3

  Because these words are appearing more frequently in non sarcastic sentences than non sarcastic sentences, classifier predicted it non sarcastic.

- Consider this Italian sentence : *#Sport - Mondiali, la disfatta mai vista del Brasile.* #non_sarcastic

- Converted to English using Machine Translation : *#Sport - World Cup defeat ever in Brazil.* Classified as Sarcastic.

  Significant words present : world, cup, defeat,ever .

  Frequency of world* words in English Sarcastic sentence : 62

  Frequency of world* words in English Non Sarcastic sentence : 40

  Frequency of cup* words in English Sarcastic sentence : 10

  Frequency of cup* words in English Non Sarcastic sentence : 3

  Frequency of defeat* words in English Sarcastic sentence : 2

  Frequency of defeat* words in English Non Sarcastic sentence : 0

  Frequency of ever words in English Sarcastic sentence : 53

  Frequency of ever words in English Non Sarcastic sentence : 34

  Because these words are appearing more frequently in sarcastic sentences than non sarcastic sentences, hence classifier predicted it sarcastic.

6.3 Conclusion

- CLSD follow CLSA in terms of comparison with In Language experiments.
- Close language give better results in CLSD as compared to languages which are not close.
- Training on Natural language is better as compared to training of translated language.
- Techniques used for Cross Language Sentiment Analysis is not adequate for Cross Lingual Sarcasm Detection.

6.4 Future Work

The field of cross lingual Sarcasm Detection has not been explored yet. There is definitely a lot of scope for improvement in this approach.

- Comparison of closeness of languages. Languages closer to each other in terms of structure will show better result as compared to other languages.
- We need better Machine Translation systems which treat sarcastic sentences differently than normal sentences.

We will be exploring this field and will try to come up with better approaches in future.

Bibliography

[1] Bing Liu. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, 2012.

[2] Tetsuya Nasukawa and Jeonghee Yi. Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture*, pages 70–77. ACM, 2003.

[3] Pushpak Bhattacharyya Balamurali, Aditya Joshi. Cross-lingual sentiment analysis for indian languages using linked wordnets. In *Proceedings of COLING 2012,COLING 2012, Mumbai*, pages 73–82, 2012.