Insult Detection in Hindi

Chetan Dalal(11218) Shivyansh Tandon(11690) Guide: Prof. Amitabha Mukherjee

Indian Institute of Technology – Kanpur, India

Problem Statement

- To detect Insults in one-liner Hindi comments in blogs and conversations.
- Given a tagged corpus of M sentences and untagged of N sentences, we use the tagged data to obtain good accuracy on the unseen to identify if the new comment is insult or not.
- Insults may be racial slurs, abuses, taunts, sarcasm, provocative words etc.

Eg. To determine if "ये चूतिया रमेश साह साला बहुत हरामी जीव है!" is an insult or not.

Present Importance

- With the increase of technology and availability of 'Hindi' language in mobiles and computers, it has become necessary to also have systems capable of detecting insults.
- The exponential growth of social networking sites, blogs etc. have encouraged us to do build such a system, because a human cannot monitor all the data that is flooded on the internet.
- For Eg: Over **409 million** people view more than **14.7 billion** pages each month and users produce about **44.5 million** new posts and **56.7 million** new comments each month. *(source: WordPress)*
- Hence, a need arises to filter all this data.

Past Works

- No work has been done for Insult detection in Hindi.
- Part of Speech n-grams, skipgrams, pattern matching approaches have been used**.**

Challenges with Dataset

- No proper dataset available.
- Used Google Translate which was not very efficient.(~1000 comments)
- Manually copied comments from various Hindi blogs.(~250 comments)

Google Translate is not that Awesome!!!!!

Word Translations

- Severity of words reduced

"Dickhead" → "गधे",

- Meaning lost

"Slut" → "फूहड़"

- Some words not translated

"Shitbag" → "Shitbag"

- Word to Word translation

"Carpet muncher" → "कालीन Muncher"

Sentence Translation

- Long sentences contain error

"Define the war on women please. You mean requiring me to pay for your birth control? That's about it. You are dead on with the talking points." → "आप अपने जन्म नियंत्रण के लिए भुगतान करने के लिए मुझे जरूरत मतलब कृपया महिलाओं पर युद्ध परिभाषित? कि इसके बारे में. आप बात कर अंक के साथ पर मर चुके हैं"

- Short sentences are better

"You're an idiot." → "तुम मूर्ख हो."

Idiom Translation

- Some idioms are fine

"Every Cloud Has A Silver Lining" → "दुर्भाग्य के काले बादलों में आशा की सुनहरी दामिनी भी छिपी रहती है"

"A Piece Of Cake" → "कोई आसान सा काम"

- Some are translated word to word and meaning is lost

"Icing On The Cake" → "केक पर टुकड़े"

"A Wolf In Sheep's Clothing" → "एक भेड़िया में भेड़ के वस्त्र"

Implementation

- We use a 4 step process.

Normalization → Feature Extraction → Feature Extraction → Classification

Normalization

- Removing random characters like '\\', '\\n' etc.
- Removing punctuations, numbers etc.
- Removing words which come only once.
- Removing the words which come very frequently.

Feature Extraction

- Sentences converted into words vector.
- Bag of Words model
- N-gram, Skip Grams
- Tf-Idf score
- Special cases of Second-Person Narrative
- Taking the special case of negation.
- For Eg: "*तुम किसी के अच्छे दोस्त नहीं बन सकते।*"

Feature Selection

- Since the number of features are large, Chi-Square test is used to find the co-relation of features with the label.
- Features with maximum value of Chi-Square statistics are selected.

Classification

- Logistic Regression and SVM models are combined to evaluate the insult.

Results

- No past results as such to compare with.
- Accuracy without applying second person feature and negation feature: **85.96**
- Accuracy with negation feature: **87.96**
- Accuracy with second-person feature: **86.96**
- Accuracy with both second-person and negation feature: **87.71**

Feature	Accuracy	Recall	Precision
Without second person and negation	85.96	0.60	.84
With Negation feature	87.96	0.67	.86
With Second-Person	86.96	0.63	.86
With both second person and negation	87.71	0.68	.84

Suggestions for Future

- To have an even larger dataset.
- To have a bigger corpus to improve accuracy.
- To identify sarcasm.
- To identify insults that have been tampered. For Eg: *"बेक-कूफ"*

References

- Sentiment Analysis For Hindi Language, MS Thesis IIIT-H 2013, Piyush Arora.Vestibulum turpis quam, tristique vel dapibus et, scelerisque luctus enim.
- Xiang, G., Hong, J., & Rose, C. P. (2012). Detecting Oensive Tweets via Topical Feature Discovery over a Large Scale Twitter Corpus, Proceedings of The 21st ACM Conference on Information and Knowledge Management, Sheraton, Maui Hawaii, Oct. 29- Nov. 2, 2012.

- For starter code:
  http://www.kaggle.com/c/detectinginsults%25E2%2580%2593in-social-commentary/forums

- For dataset:
  1. http://www.kaggle.com/c/detecting-insults-in-social-commentary/data
  2. http://ek-ziddi-dhun.blogspot.in/2008/08/blog-post_22.html
  3. http://khabar.ndtv.com/news/zara-hatke/90-per-cent-indians-are-idiots-justice-katju-357932

- Extra:
  1. http://www.wikipedia.com
  2. http://www.wordpress.com