Native Language Identification from English Writing

SNLP TERM PROJECT

UNDER THE GUIDANCE OF PROF. PAWAN GOYAL

GROUP 8

J RATAN RAHUL 10CS30022

P M ARPITA 10CS30041

CH ARCHITHA 10CS30015

YETESH CHOUDHARY 10CS30044

ASHISH VASAVA 10CS30009

PUTTY VIKAS 09CS1044

ARUNENDHER SINGHH 10CS30008

ASHWANI ATTRI 10CS30010

KONTHAM RAJESH 10CS30026

KAUSHIK KUMAR MAHATO 10CS10018

KANWAR PAL DHANDE 10CS10017

SHREYAS KHAIRKAR 11CS30034

Outline

- Objectives
- Introduction
- Dataset Construction
- Approach Taken
- Experimental Results
- Examples
- Reference

Objective

- Aim to automatically identify the native language of a writer from its English writings(articles/blogs/essays etc.)
- Applications:
  - Authorship Profiling
  - Education : more targeted feedback to language learners

Introduction

- For many years it has been presumed that the only major source for syntactic errors in adult second language performance was the performer's first language.
  - *Today in this article I would discuss about the Database Mail which is used to send the Email using SQL Server.*
  - *In today's article I will discuss Database Mail, which is used to send email using SQL Server.*
  - The first sentence can be very easily translated to an Indian language while the second cannot be translated as easily

Introduction

- Subsequent empirical studies showed that many errors are not traceable to the performer's first language but are common to second language performers of different linguistic backgrounds
  - For example, Nouns in Slavonic languages like Czech and Slovakian do not distinguish between singular and plural.
- These type of errors are found to occur more often than syntactic errors.
- This explains the necessity of supervised learning over reliable corpus dataset for Native Language Identification

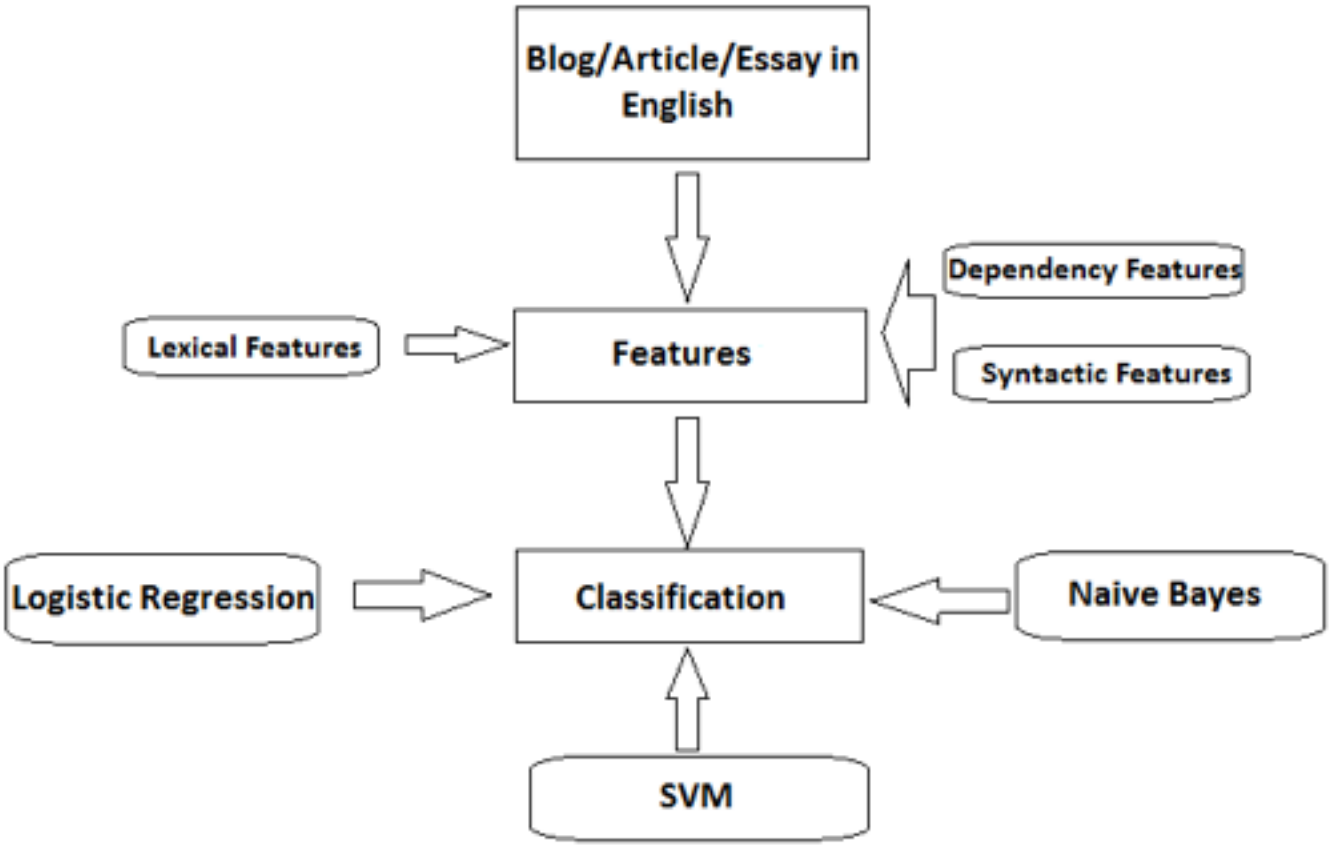
Dataset Construction

- Two datasets were used
  a. Cambridge Learners Corpus First Certificate in English (CLC-FCE)
  b. The International Corpus Network of Asian Learners of English (ICNALE)
- In total, 7765 essays from 19 languages were used with training data to test data ratio of 70:30
- More than 1000 journal entries were extracted from the popular language learner website, lang-8 for using as test data.

Dataset construction

- Languages include:
  - Catalan
  - Chinese
  - English
  - French
  - Filipino
  - German
  - Greek
  - Indonesian
  - Italian
  - Japanese
  - Korean
  - Polish
  - Portuguese
  - Russian
  - Spanish
  - Swedish
  - Thai
  - Turkish
  - Urdu

Flow Diagram



Approach Taken

- The problem is addressed as a supervised multi classification task.
- Trained our data on feature set using 4 classification models.
  - Logistic Regression
  - Gaussian Naïve Bayes
    - One vs rest classifier
  - Support Vector Machine
    - Kernel – 'rbf', tol = '0.001'
  - Multinomial Naïve Bayes
    - One vs rest classifier

Approach Taken

- The feature set comprises of the following features:
  - **Functional Words**: Certain functional words are more common in one language compared to the others. The functional words include The, to, I, and, a, was etc. 45 functional words are used as features.
  - **Word n-gram**: Writer's native language influences their choice of words. So, 1-gram and 2-gram words are used for calculating the features. 70 features are extracted from 1-gram and 290 features from 2-gram.
  - **Use of Punctuation**: The speakers of different languages use punctuation in different ways. The following 2 features are considered:
    - The number of punctuation marks used for sentence.
    - The number of punctuation marks used for word.

Approach Taken

- **Number of Unique Stems**: Speakers of different native languages differ in the amount of vocabulary used. The relative frequency of the total number of unique stems is considered as a feature.
- **Misuse of articles**: The number of instances in which an article is inconsistent with the plural and uncountable nouns are considered as features.
- **Missing Punctuation:** The relative frequency of missing punctuation after introductory (however, furthermore)and subordinating conjunction(after, before, even though) phrases are considered as features.

Approach Taken

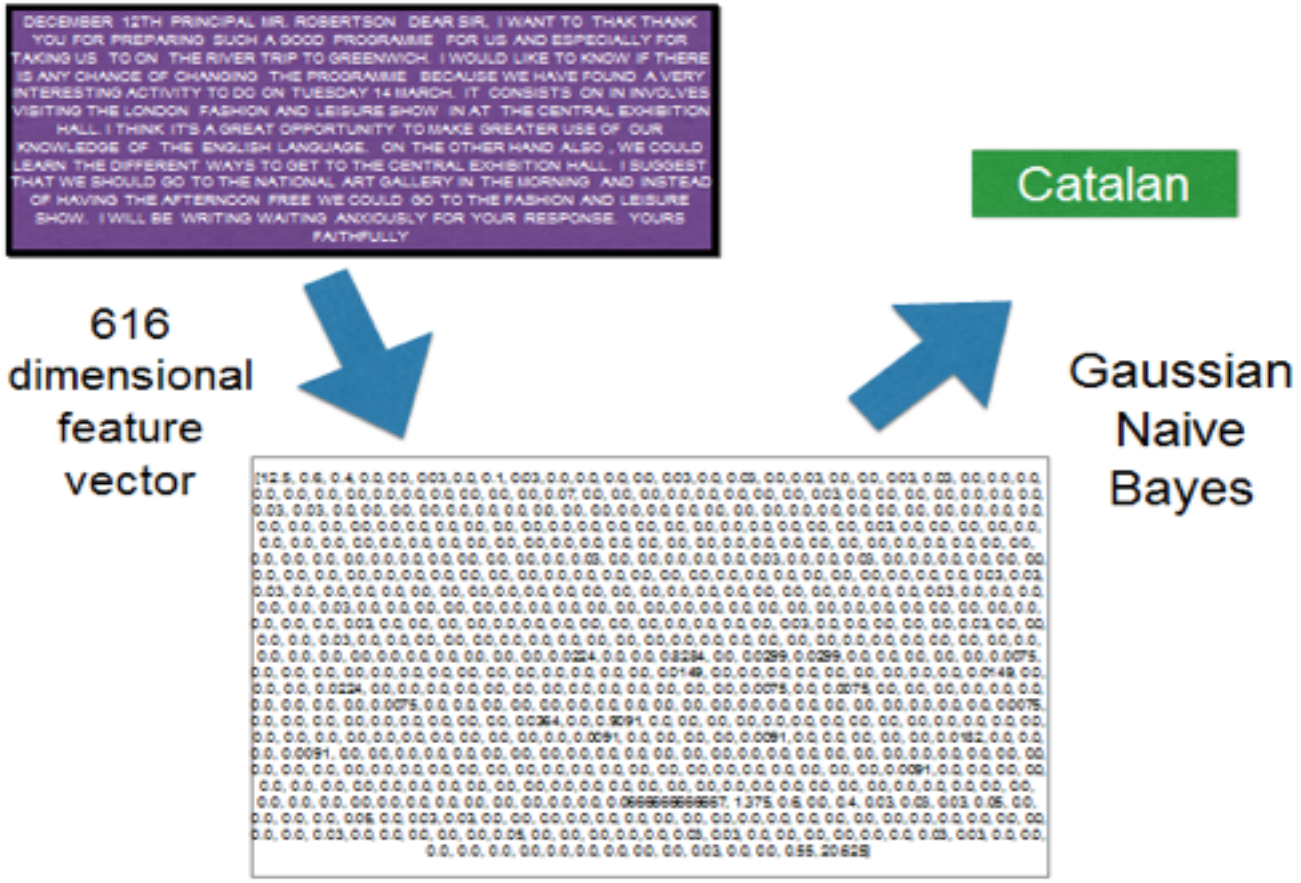
- **Words per sentence**: Number of words per sentence is considered as a feature.
- **Tense and Aspect Frequency**: Three tenses (present, past, future) are considered for calculating tense frequency. Four aspects(simple, perfect, progressive, perfect progressive) are considered for calculating aspect frequency. These frequencies are considered as features.
- **Part-of-speech**: Bigram and trigram pos tags are used for calculating features. Total of 101 bigram pos tags and 131 trigram pos tags are extracted as features.
- **Passive Constructions**: The count of number of times an author uses passive constructions (count of nsubjpass) is considered as a feature.

Results

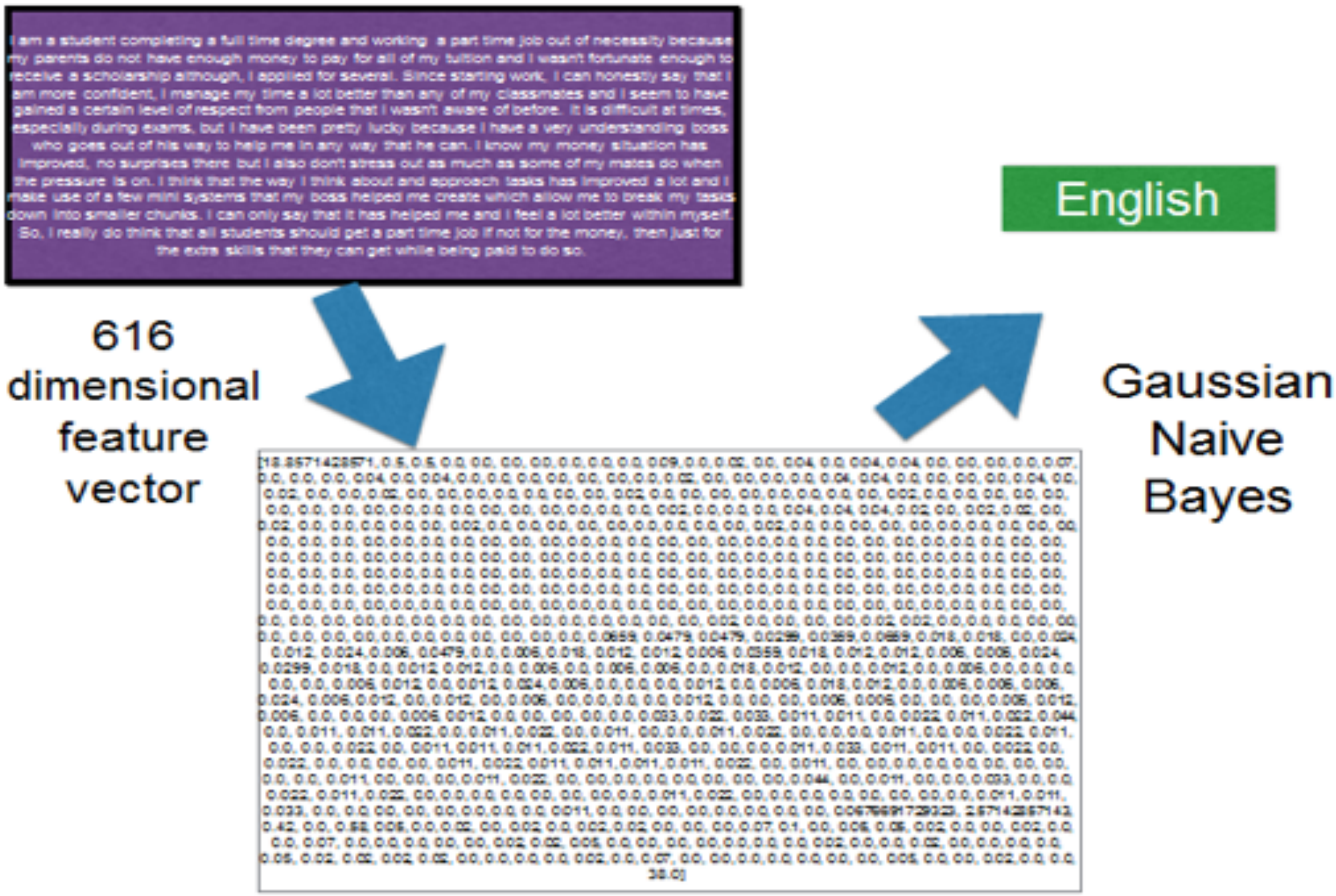
- Used a supervised multi class classifier approach for modeling the data
- Ratio of Training: Test set = 70:30
- Total of 5462 training features and 2303 test features sampled over all 19 languages.

Method	Training Accuracy	Test Accuracy
Logistic Regression	***100 %***	***55.97 %***
Gaussian Naïve Bayes	***96.33 %***	***52.36 %***
SVM	***95.97 %***	***23.75 %***
Multinomial Naïve Bayes	***59.08 %***	***13.28 %***

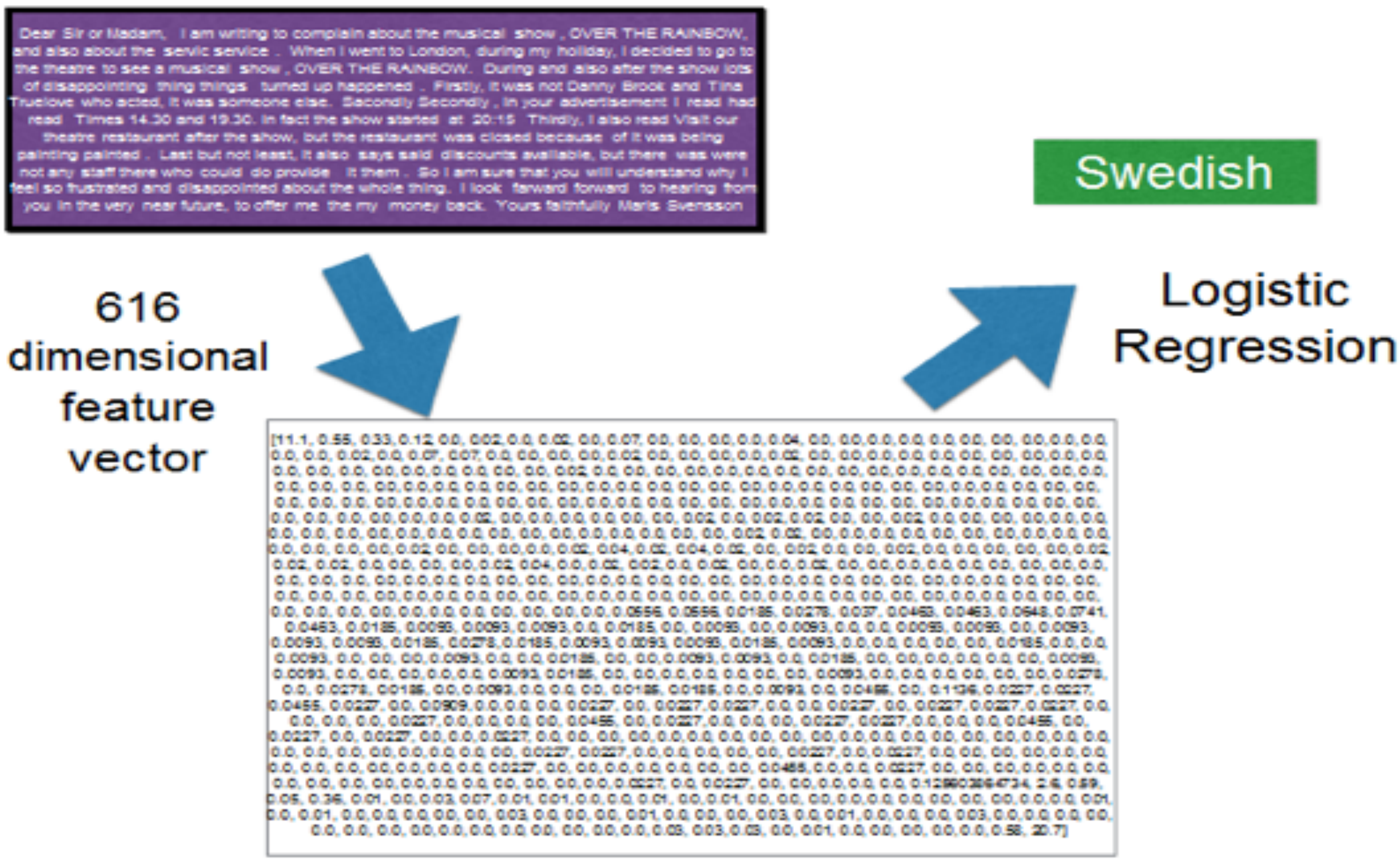
Examples


DECEMBER 12TH PRINCIPAL MR. ROBERTSON DEAR SIR, I WANT TO THAK THANK YOU FOR PREPARING SUCH A GOOD PROGRAMME FOR US AND ESPECIALLY FOR TAKING US TO ON THE RIVER TRIP TO GREENWICH. I WOULD LIKE TO KNOW IF THERE IS ANY CHANCE OF CHANGING THE PROGRAMME BECAUSE WE HAVE FOUND A VERY INTERESTING ACTIVITY TO DO ON TUESDAY 14 MARCH. IT CONSISTS ON IN INVOLVES VISITING THE LONDON FASHION AND LEISURE SHOW IN AT THE CENTRAL EXHIBITION HALL. I THINK IT'S A GREAT OPPORTUNITY TO MAKE GREATER USE OF OUR KNOWLEDGE OF THE ENGLISH LANGUAGE. ON THE OTHER HAND ALSO , WE COULD LEARN THE DIFFERENT WAYS TO GET TO THE CENTRAL EXHIBITION HALL. I SUGGEST THAT WE SHOULD GO TO THE NATIONAL ART GALLERY IN THE MORNING AND INSTEAD OF HAVING THE AFTERNOON FREE WE COULD GO TO THE FASHION AND LEISURE SHOW. I WILL BE WRITING WAITING ANXIOUSLY FOR YOUR RESPONSE. YOURS FAITHFULLY
Catalan
616 dimensional feature vector
Gaussian Naive Bayes

Examples


I am a student completing a full time degree and working a part time job out of necessity because my parents do not have enough money to pay for all of my tuition and I wasn't fortunate enough to receive a scholarship although, I applied for several. Since starting work, I can honestly say that I am more confident, I manage my time a lot better than any of my classmates and I seem to have gained a certain level of respect from people that I wasn't aware of before. It is difficult at times, especially during exams, but I have been pretty lucky because I have a very understanding boss who goes out of his way to help me in any way that he can. I know my money situation has improved, no surprises there but I also don't stress out as much as some of my mates do when the pressure is on. I think that the way I think about and approach tasks has improved a lot and I make use of a few mini systems that my boss helped me create which allow me to break my tasks down into smaller chunks. I can only say that it has helped me and I feel a lot better within myself. So, I really do think that all students should get a part time job if not for the money, then just for the extra skills that they can get while being paid to do so.
616 dimensional feature vector
Gaussian Naive Bayes
English

Examples



Dear Sir or Madam, I am writing to complain about the musical show , OVER THE RAINBOW, and also about the servic service . When I went to London, during my holiday, I decided to go to the theatre to see a musical show , OVER THE RAINBOW. During and also after the show lots of disappointing thing things turned up happened . Firstly, it was not Danny Brook and Tina Truelove who acted, it was someone else. Secondly Secondly , in your advertisement I read had read Times 14.30 and 19.30. In fact the show started at 20:15 Thirdly, I also read Visit our theatre restaurant after the show, but the restaurant was closed because of it was being painting painted . Last but not least, it also says said discounts available, but there was were not any staff there who could do provide it them . So I am sure that you will understand why I feel so frustrated and disappointed about the whole thing. I look farward forward to hearing from you in the very near future, to offer me the my money back. Yours faithfully Maria Svensson
616 dimensional feature vector
Swedish
Logistic Regression

Reference

- Abu-Jbara, Rahul Jha, Eric Morley and Dragomir Radev, Experimental results on the Native Language Identification Shared Task, *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 82–88,2013.