Understanding People in Low Resourced Languages

Thesis submitted in partial fulfillment
of the requirements for the degree of

Masters of Science in Computer Science
by
Research

by

Sahil Swami
201302071
sahil.swami@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
November 2018

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "Understanding people in Low Resourced Languages" by Sahil Swami, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date	Adviser: Prof. Manish Shrivastava

To my Friends and Family

Acknowledgments

I would like to thank my advisor, Dr. Manish Shrivastava for his guidance and expertise over these years. I would also like to thank Syed Sarfaraz Akhtar for his guidance and motivation and helping me with the research topics and suggesting how to work on them.

I'm very grateful to my parents for supporting me in everything I've decided to do. I can never thank them enough for teaching me life lessons that have always helped me in my life and for always believing in me.

I would also like to thank Shyamli for reviewing my drafts and helping me to finish them. She has always motivated me and kept me positive. I am really grateful to Mohit Agarwal for always assisting me whenever I needed his guidance in any field.

I would also like to thank my friends Ankush, Gorang, Danish, Ashutosh, Deepanshu and Aishwary for always being with me whenever I needed them and for always keeping me motivated.

Abstract

Social media platforms like Twitter and Facebook have become two of the largest platforms for people to communicate and share their views with the people. The casual and informal environment on these platforms leads to more people expressing themselves in their native language which results in a larger amount of code-mixed data that the annotated set of data currently lacks.

With access to public opinion on nearly every topic, we can gather a huge amount of user data which could prove to be useful for various companies, thus making tasks like opinion mining and sentiment analysis even more important. Hence understanding users in low resourced languages has become one of the most researched tasks of late.

We present two English-Hindi code-mixed datasets and to evaluate these datasets we simultaneously build baseline classification systems to evaluate them. As it takes time to create the datasets we decided to test our classification system on another dataset of Spanish and Catalan tweets on Catalan Independence as Catlan is one of the low resourced languages when seen from the perspective of Natural Language Processing. Thus, we first present a supervised classification system for stance and gender detection in Spanish and Catalan tweets on Catalan Independence.

Then we present two English-Hindi code-mixed corpus, one for stance detection and the other for sarcasm detection in code-mixed tweets. The tweets for stance detection are collected for the target 'Demonetisation' whereas the tweets for sarcasm detection are collected on various topics such as cricket, bollywood, and politics. Each tweet in their respective datasets is marked for the stance and presence of sarcasm. Each token in the tweets is annotated with a language tag.

Finally, we present a classification system developed using these datasets for stance and sarcasm detection. This system uses various word and character level features along with three different classification techniques. 10-fold cross-validation is used for evaluation of this system.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

One of the most spoken languages in the world is Hindi, yet if we look at it from the perspective of Natural Language Processing, it is among the lowest resourced languages. With the growth of social media platforms such as Facebook and Twitter and people expressing themselves in multiple languages, the lack of language resources makes it very difficult to perform NLP tasks to understand users and their views.

In this thesis, we aim towards working on these low resourced languages to understand users better when they express themselves in their native language on social media platforms. Understanding people broadly means understanding their sentiment, opinion and stance towards a particular target and thus it brings in the tasks of stance and sarcasm detection.

1.1 Importance of Social Media

Social media has become one of the main channels for people to communicate and share their views with the rest of the world. In recent times, social media platforms such as Facebook and Twitter, have gained a lot of popularity. These platforms offer people a medium to connect with friends, family, and colleagues, and express their opinions freely on various topics. The language used on these platforms is generally more casual and informal [17] i.e. more number of people use their native language to express themselves on these platforms. This, in turn, results in code-switching and code-mixing in texts used on social media.

1.2 Code-Mixing

Our work is on low resourced languages with more focus on English-Hindi code-mixed social media texts. Code-mixing is the conversion of one language to another within the same utterance or in the same oral or written text [18]. Code-switching and code-mixing are two of the most commonly studied phenomena in multilingual societies [32]. Code-switching is generally inter-sentential while code-mixing is intra-sentential.

With Hindi being the fourth most spoken language in the world with 41% of the Indian population speaking Hindi, and English being the lingua franca of India, English-Hindi is the most commonly used code-mixed language pair on social media. Some examples of English-Hindi code-mixed sentences:

Sentence: *modi ji notebandi ki dikkat ko door karne k liye 200 rupay ka note bi market me laao. Chae 50 ka band ho jaye.*

Words such as 'market' are in English, and words like 'ki', 'door', etc. are Hindi words which are transliterated into English.

Sentence: *Dear sir Lagta hai bina tayari ka notebandi hua hai 2000 ka note ka size kam nahi karna chahiye.*

This sentence contains words in English such as 'Dear', 'sir' and words in Hindi such as 'hai', 'bina', etc. which are transliterated to English.

1.3 Stance Detection

In this work, we mainly work on stance detection and sarcasm detection in social media texts. Stance detection is the task of automatically determining from the text whether the author is in favor or against or is neutral towards a target.

Stance detection is related to sentiment analysis but is very different from it. In sentiment analysis we check if a tweet has a positive, negative or neutral emotion while in stance detection we check whether the tweet is in favor, neutral or against a given target. For example, consider the following sentence: "*Recent studies have shown that global warming is in fact real*". We can say that this sentence's author is most likely to be in favor of the concept 'global warming'.

With the increase in the use of social media platforms by people to express their views, the task of opinion mining and sentiment analysis on natural language texts in social media has gained a lot of popularity and importance. We can find opinions on nearly every topic may it be sports, politics or movies. Researchers call this kind of data, the *Big Data*, characterized by "3V" which stands for *Volume, Variety* and *Velocity*. Some also refer to it as "5V" i.e. for *Value* and *Veracity* [15].

We can often detect from these views whether the person is in favor, against or neutral towards a given topic. There have been several experiments in the field of opinion mining on social media and online texts [20],[28]. Opinion mining can provide a lot of information about the texts present on social media and can benefit many other tasks such as information retrieval, text summarization, etc. These opinions from social media are also very useful for various companies.

We worked on stance detection in Spanish and Catalan tweets towards the target Catalan Independence. After that, we worked on stance detection in English-Hindi code-mixed tweets towards the target Demonetisation that was implemented in India in 2016.

1.4 Sarcasm Detection

The Oxford dictionary[1] defines sarcasm as: "the use of irony to mock or convey contempt". Sarcasm generally has an implied negative statement but a positive surface sentiment [19]. As an example, consider the tweet:

"*I'm so happy the teacher gave me all this homework right before Spring Break*".

The author of this tweet uses positive words like 'happy' but it can be clearly observed that the author is not happy. Although sarcasm cannot be completely formally defined, it can be detected by humans in texts and speech. Sarcasm and irony, though different, are very closely related [6], so we consider them same in our work towards sarcasm detection.

Twitter is one of the most used social media platforms used by people to express their opinions [10]. Generation of such large user data has made NLP tasks like sentiment analysis and opinion mining much more important. Many companies use this data for opinion mining and sentiment analysis to study the market. But a tweet may not always state the exact opinion of the user i.e. if it is sarcastically expressed. As it has become a common trend to use sarcasm in social media texts, detecting sarcasm in a tweet becomes more crucial and challenging for tasks like opinion mining and sentiment analysis.

The task of sarcasm detection in the text is gaining more and more importance for both commercial and security services.

We worked on sarcasm detection in English-Hindi code-mixed tweets on various subjects such as bollywood, cricket, politics, etc.

[1] http://www.oxforddictionaries.com/

1.5 Contributions of this Thesis

In this thesis, we work on understanding users in low resourced languages and thus we start by presenting a supervised classification system for stance and gender detection in Spanish and Catalan tweets directed towards Catalan Independence.

Next, to help with the lack of resources, we showcase an English-Hindi code-mixed dataset for stance detection which consists of 3545 tweets on opinion towards Demonetisation that was implemented in India in 2016. Each of the tweets is annotated for stance towards Demonetisation and each token is annotated with a language tag.

Continuing with the work on creating datasets, we present another English-Hindi code-mixed dataset for sarcasm detection which consists of tweets on various topics such as cricket, bollywood, politics, etc. where each tweet is marked for the presence of sarcasm and each token is annotated with a language tag.

Moving on from dataset creation, we then present a supervised baseline classification system for both stance and sarcasm detection in English-Hindi code-mixed tweets. This system uses various word and character level features along with three different machine learning techniques and 10-fold cross validation for evaluation.

1.6 Thesis Organisation

This thesis is divided into 7 chapters. In Chapter 2 we explain the work previously done in this field. In Chapter 3 we describe the work done on stance and gender detection in Spanish and Catalan tweets. The later two chapters describe the two English-Hindi code-mixed datasets created for stance and sarcasm detection. Chapter 6 talks about the supervised classification system developed using the datasets described in the previous two chapters. The system described in this chapter uses various machine learning models along with different word and character level features for classification.

We conclude in Chapter 7 and propose the future work that can be done.

Chapter 2

Related Work

With the increasing usage of social media and people using multilingual texts in their social media posts, code-mixing has become one of the most researched topics in Natural Language Processing. A lot of work has been done on Code-mixed social media texts. People have worked on different language pairs including English-Hindi, Arabic-Moroccon, English-Spanish, Turkish-German, etc. Researchers have presented new datasets for different languages pairs along with systems built for these datasets to perform tasks such as language identification, word normalization, etc.

Opinion mining and sarcasm detection are considered two of the major challenges to sentiment analysis. With sentiment analysis being one of the widely researched tasks in Natural Language Processing has resulted in a lot of studies on stance detection as well as sarcasm detection. People have presented new datasets for stance as well as sarcasm detection in languages other than English along with various classification approaches for detecting the same.

2.1 Code-Mixed

Various English-Hindi code-mixed datasets [32],[16] have been created for different NLP tasks. The first study is about English-Hindi text collected from Facebook forums. They also explore different NLP tasks such as language identification, normalization and POS tagging of the dataset created. Their work is focussed on POS tagging the corpus created while trying to address different challenges such as code-mixing, transliteration, non standard spelling and lack of annotated data.

The second research is about English-Hindi code-mixed dataset created by collecting texts from facebook group chats on daily life. They initially develop a language identification and word normalization system for English-Hindi code-mixed social media text. To help with the lack of annotated data for the same they create a new dataset and use the previously developed system to help with the annota-

tion of language tags and word normalization. Errors made by the system in annotation were manually corrected to make the corpus better.

2.2 Stance and Sarcasm Detection

There have been a lot of studies [28],[11] on stance detection and sentiment analysis as they are very closely related and help in various other tasks such as information retrieval and text summarization. In these studies they presented a dataset of tweets where each tweet is annotated for stance and sentiment towards specific targets. They compare different classification techniques on a dataset of Spanish tweets for sentiment analysis and topic detection.

Sarcasm detection and stance detection are both the tasks of understanding about what a person is trying to express and thus are very similar to each other. A lot of researches [3],[6],[1],[30], [22],[5] have been performed on sarcasm detection in various different languages such as English, Czech, Dutch and Italian. One of the work explores various lexical and pragmatic based features where one of the other puts emphasis on the importance of pattern-based features for classification. They also compare various supervised and semi-supervised classification techniques for sarcasm detection in social media texts. Some of the studies presented new datasets for sarcasm detection in languages other than English and presented language independent classification systems and compared it with sarcasm detection in an English dataset.

Chapter 3

Stance and Gender detection in Spanish and Catalan tweets

Catalan being the second most spoken language in Spain, is a very low resourced language when considered from the perspective of Natural Language Processing tasks. To work on a new Catalan and Spanish dataset we decided to take part in the task of stance and gender detection in Spanish and Catalan tweets organized by IBEREVAL. They provided a dataset of Spanish and Catalan tweets marked for stance towards Catalan Independence. In this work, our main aim is stance detection in low resourced languages, and therefore this task makes it perfect for us to participate in it as Catalan is a low resourced language.

In this chapter, we describe the system submitted to IBEREVAL-2017 for stance and gender detection in Spanish and Catalan tweets on Catalan Independence [24]. We developed a supervised system using Support Vector Machines with radial basis function kernel to identify the stance and gender of the tweeter using various character level and word level features. Our system achieves a macro-average of F-score(FAVOR) and F-score(AGAINST) of 0.46 for stance detection in both Spanish and Catalan and an accuracy of 64.85% and 44.59% for Gender detection in Spanish and Catalan respectively.

3.1 Introduction

As mentioned in previous chapters, there have been several experiments in the field of sentiment analysis and opinion mining on social media texts it can provide a lot of information about the texts that are present in social media and benefits a lot of other NLP tasks.

On the other hand gender detection is the task of inferring the gender of the author from the content of the tweet. Gender detection has many applications in the field of marketing and advertising and thus there have been a lot of studies [8, 27, 29, 7] on gender detection in social media text. Twitter profiles don't provide a field for persons gender which makes the task of identifying author's gender from the tweet much more important.

3.2 Dataset and Evaluation

The organizers provided training and test dataset which consisted of 4319 tweets and 1081 tweets for both Spanish and Catalan respectively. All the tweets in the training dataset are annotated with stance (FAVOR or AGAINST or NONE) and gender (FEMALE or MALE). Here are some examples from the dataset:

Tweet id: *54e6b766931cd6722314cad0cbc2ad8e*
Tweet: *Tuits Tsunami! Optimistic about the future? #Elecciones #ComunicacinPoltica #VamosJuntos #LlamadasQueUnen #CaminemosJuntos #Cambiemos #27S*
Stance tag: *AGAINST*
Gender tag: *FEMALE*

Tweet id: *cace4e761867edff088f34786a7b103f*
Tweet: *Pues no, Independencia si o si, y he votado a la CUP #eleccionescatalanas*
Stance tag: *FAVOR*
Gender tag: *MALE*

We were asked to submit a maximum of five runs that contained the stance tags and gender tags along with the tweet id for the test data and then our systems were evaluated using those tags.

Stance detection systems were evaluated using macro-average of F-score (FAVOR) and F-score (AGAINST) i.e.

$$(Fscore_{FAVOR} + Fscore_{AGAINST})/2$$

On the other hand, gender detection systems were evaluated using accuracy i.e. the number of tweets for which the gender is predicted correctly per hundred tweets.

3.3 System Framework

In this section, we describe the features and classification technique used in this system. We also describe the processing of data before extracting the features and the feature selection technique used to reduce the feature vector size.

3.3.1 Pre-processing

Initially, tweets are tokenized in a way such that hashtags, URLs, and mentions are preserved. Then URLs, mentions, and stopwords are removed from the tweets.

It can be observed from the tweets present in the training and test datasets that almost all the hashtags are written in camel case format. Therefore, # is removed from the hashtags and all the words are extracted from the hashtag. And then each word is considered as a separate token.

All the tokens in Spanish are then stemmed using Snowballstemmer implemented in NLTK.

3.3.2 Features

We extracted various features from the given tweets to train our machine learning model. We list and describe these features below.

3.3.2.1 Character N-grams

Character n-grams feature refers to presence or absence of a contiguous sequence of n characters. It can be seen from previous work [28, 8, 27] that character level features have a significant effect on stance and gender detection.

We extract character n-grams for all values of n between 1 and 3. Including all the n-grams increases the size of feature vector enormously. Therefore, we consider only those n-grams in our feature vector which occur at least 10 times in the training dataset. This reduces the size of feature vector significantly and also removes noisy n-grams.

3.3.2.2 Word N-grams

Word n-grams feature refer to presence or absence of a contiguous sequence of n words or tokens. Word n-grams have proven to be important features for stance and gender detection in previous studies [21, 29]. We extract word n-grams for all values of n between 1 and 5. We include only those n-grams in our feature vector which occur at least 10 times in the training dataset.

3.3.2.3 Stance and Gender Indicative Tokens

This feature refers to presence or absence of stance and gender indicative tokens. We use a variation of the approach to find stance indicative hashtags [28] and extract stance and gender indicative tokens.

We calculate a score for each token for both stance and gender where score is defined as :

$$Score_{stance}(token) = max_{stance\_label \in Stance-Set} \frac{freq(token, stance\_label)}{freq(token)}$$

$$Score_{gender}(token) = max_{gender\_label \in Gender-Set} \frac{freq(token, gender\_label)}{freq(token)}$$

where Stance-Set = {FAVOR, AGAINST, NEUTRAL}, Gender-Set = {MALE, FEMALE}.

We consider only those tokens as features for stance indication which have a score ≥ 0.6 and occur at least five times in the training dataset. For gender indication, we consider only those tokens which have a score ≥ 0.7 and occur at least twice in the training dataset. The threshold value for scores and number of occurrences has been decided after empirical fine tuning.

3.3.3 Feature Selection

Previous studies [27, 23] have shown that feature selection algorithms improve efficiency and accuracy of classification systems. It reduces the feature vector size by removing the features that have a low impact on classification. We used chi square feature selection algorithm which uses chi-squared statistic to evaluate individual feature with respect to each class. This algorithm was run for both stance and gender detection in order to extract the best features and reduce the feature vector size.

3.3.4 Classification Approach

Support Vector Machines have been used many times previously [28, 12, 26] for stance and gender detection and has proven to be a very effective classification technique for the same.

After pre-processing the dataset and extracting all the desired features, we use scikit-learn Support Vector Machine implementation with a radial basis function kernel for classification. We also perform 10-fold cross validation on the provided training dataset to develop the system. 10-fold cross validation is run for each of the individual features separately to observe the effect of each feature on classification.

3.3.5 Results

To develop and evaluate our supervised classification system we ran 10-fold cross-validation on the training dataset and calculated the accuracies for both stance and gender detection.

Table 3.1 and Table 3.2 show the accuracy in percentage achieved for stance detection and gender detection respectively for Spanish tweets while Table 3.3 and Table 3.4 show the accuracy achieved for

	Stance Detection
Character N-grams	74.94
Word N-grams	74.03
Stance and gender indicative tokens	75.40
All features	75.81

Table 3.1 Feature-Wise Accuracy (in %) for Stance Detection in Spanish Tweets.

	Gender Detection
Character N-grams	69.18
Word N-grams	63.38
Stance and gender indicative tokens	63.43
All features	69.83

Table 3.2 Feature-Wise Accuracy (in %) for Gender Detection in Spanish Tweets.

stance and gender detection for Catalan tweets considering one feature at a time and also considering all the features together. It can be observed from the results of 10-fold cross-validation on training dataset that character n-grams have a significant effect on classification.

Our system achieved a macro-average of F-score(FAVOR) and F-score(AGAINST) of 0.46 for stance detection in both Spanish and Catalan and an accuracy of 64.85% and 44.59% for gender detection in Spanish and Catalan respectively for the given test dataset. This data was provided by the organizers after evaluating our submitted runs.

	Stance Detection
Character N-grams	81.16
Word N-grams	79.48
Stance and gender indicative tokens	80.64
All features	81.53

Table 3.3 Feature-Wise Accuracy (in %) for Stance Detection in Catalan Tweets.

	Gender Detection
Character N-grams	73.64
Word N-grams	69.60
Stance and gender indicative tokens	71.34
All features	75.38

Table 3.4 Feature-Wise Accuracy (in %) for Gender Detection in Catalan Tweets.

3.4 Conclusion

In this chapter, we described our work on social media texts in two languages i.e. Spanish and Catalan that was written towards Catalan Independence on which we performed stance and gender detection by developing a supervised classification system using character and word level features and Support Vector Machine technique for classification. In the next chapter, we present our work on another low resourced domain i.e. stance detection in English-Hindi code-mixed social media text.

Chapter 4

An English-Hindi Code-Mixed Corpus for Stance Detection

After working on stance and gender detection in Spanish and Catalan tweets we decided to proceed with stance detection but in a different low resourced language. As the lack of corpus and resources poses a lot of challenges in various NLP tasks we decided to build a dataset to help with these challenges.

With Hindi being the most spoken language in India and the fourth most spoken in the world, and English being the third most spoken language in the world, a lot of people express themselves on social media in code-mixed and code-switched texts. With very few English-Hindi code-mixed datasets available, it makes it very difficult to perform NLP tasks such as stance detection on these social media texts.

In this chapter, we present the first English-Hindi code-mixed dataset for stance detection. This dataset consists of English-Hindi code-mixed tweets towards Demonetisation that was implemented in India in 2016. In chapter-5 we also present a supervised classification system for stance detection developed using the same dataset.

4.1 Introduction

As mentioned in Chapter 2, several code-mixed datasets have been created for various NLP tasks but no opinion mining experiment has been performed on English-Hindi code-mixed data. Therefore, we aim to provide an English-Hindi code-mixed dataset and perform an experiment of opinion mining on it.

We present a new dataset that consists of 3545 English-Hindi code-mixed tweets with opinion towards the target Demonetisation that was implemented in India in 2016 which was followed by a large countrywide debate.

The target for tweets in this dataset i.e. 'Notebandi' or 'Demonetisation' was implemented in India on 8th November, 2016 in which currency in the denominations of 500 and 1000 was declared invalid. The government claimed that this decision was taken to eliminate the use of counterfeit cash used to fund illegal activities and terrorism. People all over India had different reactions to this event and many of them used Twitter to express their views. Consider the following tweet: '*Demonetisation has caused a lot of problems for everyone*'. We can say that the author if this tweet is most likely to be against the target demonetisation.

This chapter describes a dataset of English-Hindi code-mixed tweets on 'Notebandi' or 'Demonetisation' with tweet level annotation for stance towards this target and token level language annotation that can be used to develop and evaluate the performance of stance detection and language identification techniques on a code-mixed corpus.

This dataset has been made available online[1].

4.2 Dataset

This section explains the process of data collection as well as the processing of data that has to be done to proceed with annotation. We also explain the process of tweet level stance annotation and token level language annotation.

4.2.1 Data Collection

We collect tweets related to the Demonetisation that was implemented in India in 2016. We use Twitter Scraper API to collect tweets using the keywords notebandi and demonetisation over a period of 6 months after Demonetisation was implemented. All the tweets that are written exclusively in English or Hindi are eliminated and code-mixed tweets are selected manually. Each tweet is collected in json format after which the content of the tweet and the tweet id are extracted from it. A total of 3545 English-Hindi code-mixed tweets are collected. Here is an example of a tweet collected in json format:

{*"timestamp": "2017-08-23T09:43:28", "text": "to aapke anusaar baal vivaah, sati pratha, vidhwa vivaah, triple talaq, halala jaise issue koi issue hi nahi hain is samaaj ke liye?", "user": "vineetdw", "retweets": "0", "id": "900292394758807552", "likes": "1"*}

[1] https://github.com/sahilswami96/StanceDetection_CodeMixed

4.2.2 Data Processing and Annotation

The tweets are annotated by a group of native Hindi speakers who are also fluent in English. Each tweet is annotated for stance towards demonetisation. Tweets are then tokenized for language annotation after which the tokenization and language tags are manually reviewed to resolve any errors. The inter-annotator agreement i.e. Cohen's Kappa on the annotations for stance [13] turned out to be 0.82. The disagreement was resolved by asking the annotators to agree on a single annotation. If the annotators were not able to agree on a particular tag, then that tweet was removed from the dataset.

4.2.2.1 Stance Annotation

Each of the tweets is manually annotated with one of the following stance tags: 'FAVOR, 'AGAINST and 'NONE. Some hashtags and keywords, such as #IAmWithModi, #ByeByeBlackMoney and 'samarthan are direct indicators that the author is in favor of demonetisation. Similarly, hashtags such as #StopDemonetisation, #NoteNahiPMBadlo, and #ModiSurgicalStrikeOnCommonMan are clear indicators that the author is against demonetisation. Examples of tweets (with translation in English) with different stances towards the target are:

Target: Demonetisation

Tweet: @narendramodi thanks for notebandi hum aap ke saath hai

Translation: @narendramodi thanks for notebandi we are with you

Stance: FAVOR

Tweet: @PMOIndia Chalo Modi ji apne Deshwasi sang majak kar liya, ab log bahut paresan hai 500/1000 pr rahem kr Notebandi wapos lo

Translation: @PMOIndia Modi ji you played a prank with the people of your country, people are really hassled. Show mercy on 500/1000 and take demonetization back

Stance: AGAINST

Tweet: Neta samajh. Nahi pa rahe hai ki notebandi par hindu muslim rajniti kaise kare , ye hi hai sabka sath sabka vikas

Translation: Political leaders are confused on how to do hindu muslim politics on demonetization, this is everyone's unity everyone's progress

Stance: NONE

4.2.2.2 Tokenization and Language Annotation

Several experiments have been performed for language identification [2],[9],[25] on monolingual and code-mixed texts which motivates the task of token level language annotation in the presented corpus.

The text written on Twitter by users is sometimes a lot different from normal texts found in documents. It is a common trend to use multiple punctuations and white spaces such as ..., ,,,, !!!, etc. It is also common to use multiple mentions, hashtags, and URLs in a tweet. We tokenize the tweets after taking this information into account and by using white spaces as delimiters. Tokenization is manually verified by multiple people proficient in both English and Hindi to correct any mistakes.

Each token is then annotated with one of the language tags: en, hi, rest. En refers to English and is assigned to English words such as 'happy', 'today', etc. hi refers to Hindi and is assigned to Hindi words transliterated in English such as 'nahi' (no), 'samajh' (understand). A token is annotated with rest when it is a named entity, punctuation, hashtag, URL or a mention, etc. Initially the tokens are automatically annotated with language tags using online available dictionaries such as 'Enchant' and the 'rest' tag is assigned by identifying hashtags, URLs, mentions and emoticons. We also create a list of popular named entities related to Demonetization to annotate named entities. Then each tag is manually verified to correct any wrong annotation. Table 4.1 shows an example of a language annotated tweet.

4.2.3 Dataset Analysis

The dataset consists of 3545 English-Hindi code-mixed tweets where each of them is annotated with stance towards Demonetisation. Each tweet is tokenized and each token is annotated with a language tag. The dataset has 964 tweets in favor, 647 tweets against and 1934 tweets that have no stance towards the target.

The average length of a tweet is 21.3 tokens per tweet. There are an average of 16.3, 2.0 and 3.0 'hi', 'en' and 'rest' tokens respectively per tweet. Figure 4.2 shows corpus level statistics whereas Figure 4.3 shows tweet level statistics. This corpus can be used for developing and evaluating opinion mining and language identification techniques.

4.2.4 Dataset Structure

The corpus is structured into three files. The first file contains a tweet id followed by the corresponding tweet text and a blank line and so on. The second file consists of tweet ids followed by language annotated tweets. The third file has the stance for each tweet. Each tweet id is followed by one of the stance tags and a blank line.

Token	Language
#Notebandi	rest
ka	hi
niyam	hi
:	rest
khata	hi
nahi	hi
hai	hi
to	hi
khulwao	hi
.	rest
Aam	hi
aadmi	hi
:	rest
khulwa	hi
to	hi
lun	hi
.	rest
Par	hi
bhai	hi
bank	en
main	hi
ghusub	hi
Kasey	hi
?	rest

Table 4.1 A Tweet with Token Level Language Annotation

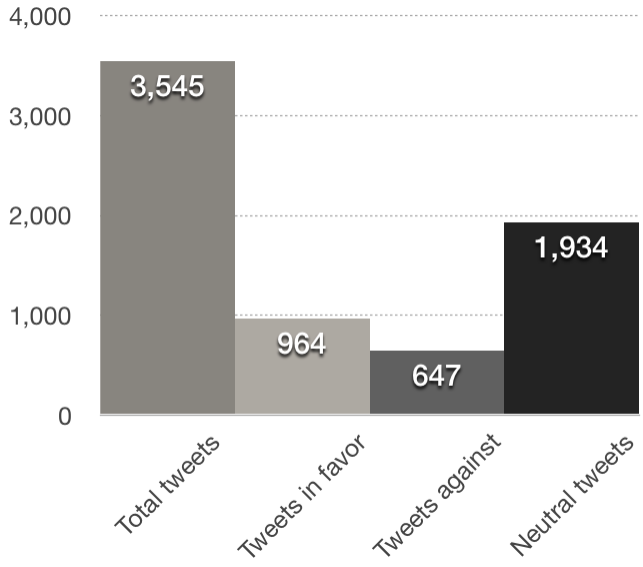

Figure 4.1 Corpus Level Statistics

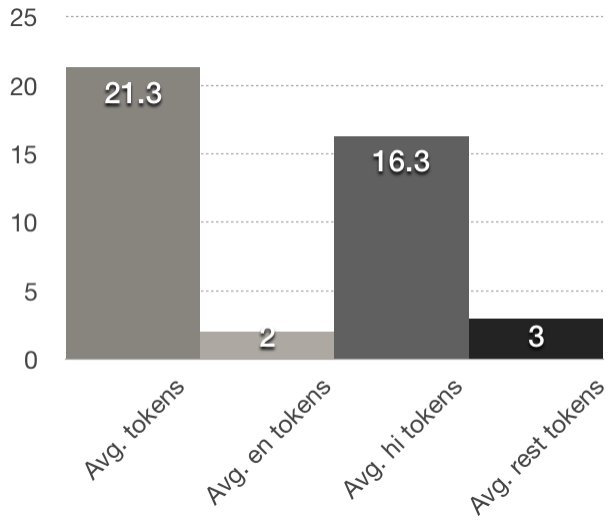

Figure 4.2 Tweet Level Statistics

4.3 Conclusion

In this chapter, we presented the first English-Hindi code-mixed dataset collected from twitter for stance identification towards Demonetisation. We explained the methods used for annotating each tweet with stance towards the target and for annotating each token with a language tag. We will present a framework for stance detection developed using the same dataset which uses three different machine learning techniques in chapter 5. These techniques are then evaluated by running 10-fold cross-validation.

Chapter 5

An English-Hindi Code-Mixed Corpus for Sarcasm Detection

After creating an English-Hindi code-mixed dataset for stance detection we decided to take our work forward in the direction of code-mixed data. With even less work done on sarcasm detection than code-mixed data, we decided to work on sarcasm detection in English-Hindi code-mixed data.

In this chapter, we present the first English-Hindi code-mixed corpus created for sarcasm detection in tweets. This dataset consists of tweets on various topics such as cricket, politics, bollywood, etc. out of which some tweets are sarcastic while the others are not. This dataset is further used to develop a supervised classification system for sarcasm detection in English-Hindi code-mixed tweets which is described in chapter 5.

5.1 Introduction

As mentioned in Chapter 2 there have been a lot of studies on sarcasm detection in various different languages but there have been no experiments on English-Hindi code-mixed texts mainly because of the lack of annotated resources. This creates a lot of challenges to perform other NLP tasks on English-Hindi code-mixed texts that can benefit from sarcasm detection.

To help with these challenges we aim to provide a dataset for the same. Thus the main contribution of this chapter is to provide a resource of English-Hindi code-mixed tweets which contain both sarcastic and non-sarcastic tweets. We provide tweet level annotation for the presence of sarcasm and token level language annotation. This corpus can be used to train, develop and also evaluate the performances of sarcasm detection and language identification techniques on a code-mixed corpus.

This dataset is freely available online[1].

[1] https://github.com/sahilswami96/SarcasmDetection_CodeMixed

5.2 Dataset

This section explains the methods used for the collection of tweets and the processing data on the tweets for feature extraction. This section also describes the method used for annotation of sarcasm in tweets and the annotation of tokens with language tags.

5.2.1 Data Collection

To collect sarcastic tweets we extract tweets containing hashtags #sarcasm and #irony [14] using the Twitter Scraper API and manually select English-Hindi code-mixed tweets from them. We also use other keywords such as 'bollywood', 'cricket' and 'politics' to collect sarcastic tweets from these domains. Out of these collected tweets, sarcastic and non-sarcastic tweets are further manually separated.

To collect more non-sarcastic tweets we extract tweets with keywords such as 'bollywood', 'cricket' and 'politics' which do not contain hashtags #sarcasm and #irony, and English-Hindi code-mixed tweets are manually selected from them. Having only sarcastic or only non-sarcastic tweets from a particular domain may lead to a biased classification system, therefore, we make sure that there are both sarcastic and non-sarcastic tweets from each domain. The twitter scraper API collects each tweet in json format after which we extract the tweet content and tweet id from it. Figure 1. shows an example of a tweet collected in json format.

5.2.2 Data Processing and Annotation

Tweets are annotated by a group of people fluent in both English and Hindi. Each tweet is manually annotated for the presence of sarcasm. Tweets are then tokenized and each token is annotated with a language which is manually verified. We used Cohen's Kappa [13] as a measure of inter-annotator agreement and it was calculated to be 0.79. The disagreement was resolved by asking the annotators to agree on a single annotation. If the annotators were not able to agree on a particular tag, then that tweet was removed from the dataset.

5.2.2.1 Sarcasm Annotation

Each tweet is manually annotated for the presence of sarcasm using the tags 'YES' and 'NO'. Tweets with the hashtags #sarcasm and #irony are more likely to contain sarcasm. Tweets which do not contain these hashtags are then manually verified to not contain sarcasm. An example of a tweet (with translation in English) that contains sarcasm and one that does not:

Tweet: *@bonda0123 sir g .. #insomniac likhte ho aur jaldi sone ki baat bhi karte ho !! #irony !!*
Translation: *@bonda0123 sir You write #insomniac and talk about sleeping early !! #irony !!*
Sarcasm: YES

Tweet: *Bhai kuchh bhi karna iss @SimplySajidK ke saath movie mat karna..Bollywood se nafrat ho jaati hai..Itni sadi hui ghatiya filmein banata h ye*
Translation: *Brother do anything but don't do a movie with @SimplySajidk..I start hating Bollywood..They make such bad films*
Sarcasm: NO

Hashtags #sarcasm and #irony are randomly removed from some tweets which contain sarcasm so that the dataset contains both types of sarcastic and ironic tweets, ones with the hashtags #sarcasm and #irony and ones without.

5.2.2.2 Tokenization and Language Annotation

There have been several experiments of language identification [2],[9] on various types of texts which motivates the task of token level language annotation in this dataset.

Each tweet is tokenized using white spaces as delimiters and taking into account the trends found in the dataset such as the use of multiple consecutive punctuations, mentions, etc. Each token is annotated with a language tag. One of the following tags is assigned for language: 'en', 'hi' and 'rest', where 'en' stands for English, 'hi' for Hindi and 'rest' for punctuations, emoticons, named entities, URLs, etc. 'en' is assigned to English words such as 'play', 'warm', etc. and 'hi' is assigned to Hindi words transliterated in English such as 'sahi', 'kya'.

Initially each token is assigned language tags using online dictionaries such as Enchant and the 'rest' tags are assigned by identifying hashtags, URLs and mentions. Every language tag and token is manually verified to correct any mistakes. Table 5.1 is an example of a tweet with language tags:

5.2.3 Dataset Analysis

The dataset consists of 5250 English-Hindi code-mixed tweets out of which 504 tweets are marked as sarcastic and ironic. The dataset consists of two types of tweets: 1.) Tweets that are marked as sarcastic but do not have hashtags #sarcasm or #irony present in them. 2.) Tweets that contain these hashtags but are not marked as sarcastic. This sparsity in the corpus also helps in developing a better system for sarcasm detection.

Token	Language
bhai	hi
triple	en
talaq	hi
se	hi
aap	hi
kya	hi
samjhte	hi
hai	hi
samjhaye	hi
aap	hi
zara	hi
..	rest
agar	hi
triple	en
talaq	hi
pta	hi
hota	hi
apko	hi
toh	hi
aisa	hi
nhi	en
kehte	hi
..	rest

Table 5.1 A Sample Tweet with Tokens Annotated for Language

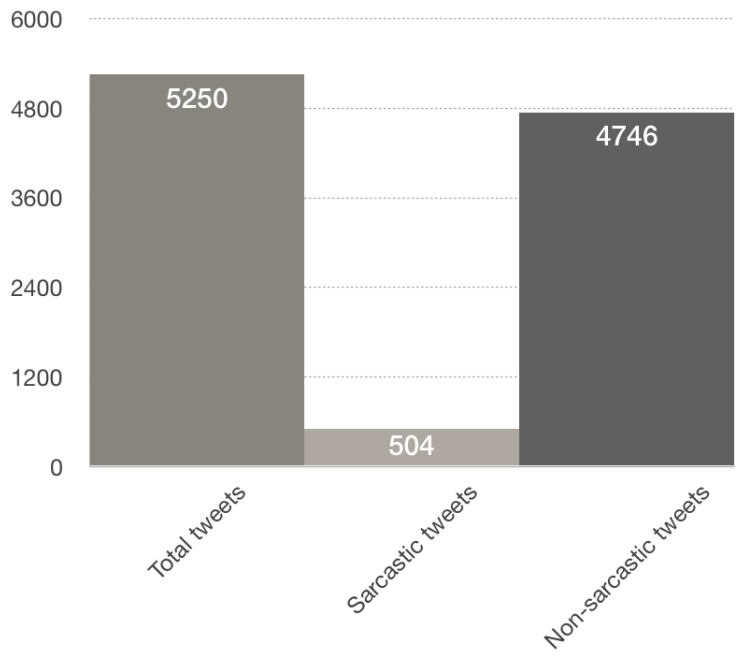

Figure 5.1 Corpus Level Statistics

The average length of a tweet is 22.2 tokens per tweet. The average number of tokens per tweet annotated with 'en', 'hi' and 'rest' tags are 2.1, 16.1 and 4.0 respectively. Figure 5.1 and Figure 5.2 show corpus level and tweet level statistics respectively.

As the number of sarcastic tweets is significantly less than the number of non-sarcastic tweets, thus when performing sarcasm detection on this dataset (described in Chapter 6), we use F-score measure for evaluation.

5.2.4 Dataset Structure

The corpus is structured into three files. The first file contains a tweet id followed by the corresponding tweet text and a blank line and so on. The second file consists of tweet ids followed by language annotated tweets as depicted in Table 1. The third file has the annotation for the presence of sarcasm for each tweet. Each tweet id is followed by one of the sarcasm label, a blank line.

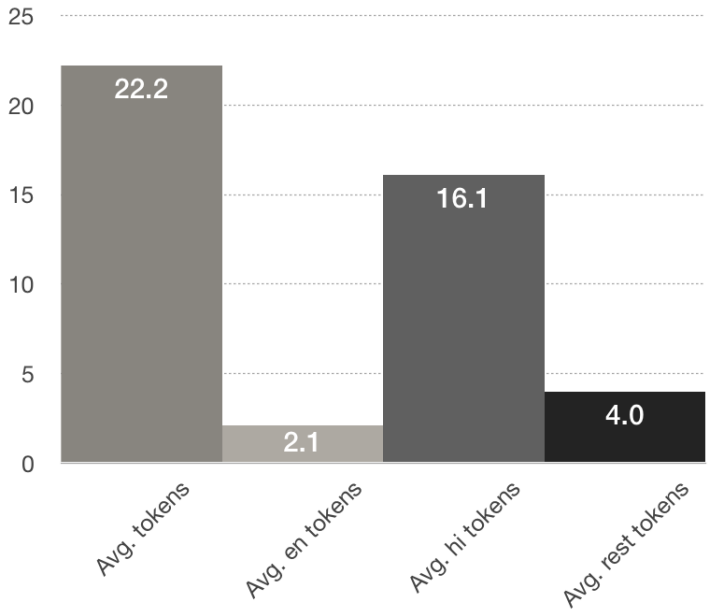

Figure 5.2 Tweet Level Statistics

5.3 Conclusion

In this chapter, we presented the first English-Hindi code-mixed dataset for sarcasm detection collected from twitter. We explained the methods used for collecting and annotating these tweets at both tweet level for presence of sarcasm as well as at token level for language.

In the next chapter, we present a supervised classification system for sarcasm detection developed using the same dataset that uses various machine learning techniques along with word and character level features. This system is then evaluated using 10-fold cross-validation.

Chapter 6

Baseline classification systems for Stance and Sarcasm detection in English-Hindi code-mixed tweets

After creating the English-Hindi code-mixed corpus for stance and sarcasm detection we developed baseline classification systems using these datasets for both stance and sarcasm detection to evaluate these datasets. In this chapter we describe the working, features used and results achieved by both the classification systems.

6.1 Classification System

The baseline classification system that we present for stance detection and sarcasm detection in English-Hindi code-mixed tweets use various character and word level features. We run various machine learning models over these features for stance detection and sarcasm detection. This classification system is available online[1]

6.1.1 Preprocessing

URLs, mentions and stop words are removed from the tweets for further processing. Hashtags are extracted for each tweet and as it is a general trend to use camel case format while writing hashtags, we remove the '#' from the hashtags and use an approach [4] for hashtag decomposition to extract all the words from the hashtag. For example, #IAmWithModi can be decomposed into four separate words i.e. 'I', 'Am', 'With' and 'Modi'. Each of these words is then treated as a separate token.

[1] https://github.com/sahilswami96/

6.1.2 Features

Here are the various word and character level features extracted from the tweets for classification:

6.1.2.1 Character N-grams

Character n-gram refers to presence or absence of a contiguous sequence of n characters in the tweet. It can be seen from previous works [28],[31] that character level features have a significant effect on stance detection. Character n-grams have proved to be one of the most important features in previous experiments [6],[30] on sarcasm detection.

We extract character n-grams for all values of n between 1 and 3. Including all the n-grams increases the size of feature vector enormously. Therefore, we consider only those n-grams in our feature vector which occur at least 8 times in the dataset. This reduces the size of feature vector significantly and also removes noisy n-grams.

6.1.2.2 Word N-grams

Word n-gram refers to presence or absence of a contiguous sequence of n words or tokens in the tweet. Word n-grams have proven to be important features for stance detection in previous studies [20],[28],[31]. Word n-grams have proven to be useful features for sarcasm detection as well in previous experiments [1],[6],[30]. We extract word n-grams for all values of n between 1 and 5. We include only those n-grams in our feature vector which occur at least 10 times in the dataset.

6.1.2.3 Stance Indicative Tokens

This feature refers to the presence or absence of stance indicative tokens. We use a variation of the approach to find stance indicative hashtags [28] and extract stance indicative tokens for each language label. We calculate a score for each token for stance where score is defined as :

$$Score(token) = max_{label \in Stance-Set} \frac{freq(token, stance\_label)}{freq(token)}$$

where Stance-Set = {FAVOR, AGAINST, NONE}.

We consider only those tokens as features for stance indication which have a score ≥ 0.6 and occur at least five times in the dataset. We find such tokens for each of the language tags and consider them

in the feature vector. The threshold value for scores and number of occurrences has been decided after empirical fine tuning.

6.1.2.4 Sarcasm Indicative Tokens

This feature refers to the presence or absence of sarcasm indicative tokens. We use the same approach as used in the previous section to extract sarcasm indicative tokens. We calculate a score for each token where the score is defined as:

$$Score(token) = max_{label \in Sarcasm-Set} \frac{freq(token, sarcasm\_label)}{freq(token)}$$

where Sarcasm-Set = {YES, NO}.

We consider only those tokens as features for sarcasm indication which have a score ≥ 0.6 and occur at least five times in the dataset. We find such tokens for each of the language tags and consider them in the feature vector. The threshold value for scores and number of occurrences has been decided after empirical fine tuning.

6.1.2.5 Emoticons

This feature refers to the presence or absence of various emoticons in the tweet. There have been several experiments [3],[30] where emoticons are used as a feature for sarcasm detection. We consider a set of 27 emoticons as features. We use this feature only for sarcasm detection as it did not have a positive impact on stance detection.

6.1.3 Feature Selection

Previous studies [23],[31] have shown that feature selection algorithms improve efficiency and accuracy of classification systems. We used chi square feature selection algorithm which uses chi-squared statistic to evaluate individual feature with respect to each class. This algorithm was used in order to extract the best features and reduce the size of feature vectors to 500.

6.1.4 Classification Approach

We compare various machine learning models using the same features for stance and sarcasm detection.

Three classification techniques have been used for this experiment:
1. Support Vector Machine with Radial Basis Function kernel
2. Random Forest classifier
3. Linear support vector machine
We use the scikit-learn implementation of these methods.

After pre-processing the dataset and extracting all the desired features, we run the above mentioned techniques and perform 10-fold cross-validation. 10-fold cross validation is run for each of the individual features separately to observe the effect of each feature on classification.

6.1.5 Results

Table 6.1, 6.2 and 6.3 show the f-score achieved in stance detection when considering a single feature at a time as well as considering all at the same time for each of the machine learning techniques. It can be observed that Support Vector Machine with Radial Basis Function kernel performs the best for stance detection. It can also be observed that all of the three features have nearly the same effect on classification.

Features	**RBF Kernel SVM**
Character n-grams	68.6
Word n-grams	66.5
Stance indicative tokens	66.9
All features	69.7

Table 6.1 F-scores for RBF Kernel SVM Classifier for Stance Detection

Features	**Random Forest**
Character n-grams	63.4
Word n-grams	62.9
Stance indicative tokens	63.6
All features	65.9

Table 6.2 F-scores for Random Forest Classifier for Stance Detection

For sarcasm detection also, we use the F-score measure to evaluate the performance of our system as the number of sarcastic tweets is less than the number of non-sarcastic tweets and thus using just accuracy for evaluation of the system may not be a good metric. F-score is defined as the harmonic mean of precision and recall.

Features	Linear SVM
Character n-grams	66.9
Word n-grams	66.1
Stance indicative tokens	66.4
All features	67.7

Table 6.3 F-scores for Linear SVM Classifier for Stance Detection

$$F - score = 2\frac{precision * recall}{precision + recall}$$

Precision and recall are defined as:

$$Precision = \frac{tp}{tp + fp}$$

$$Recall = \frac{tp}{tp + fn}$$

where *tp*, *fp* and *fn* are true positives, false positives and false negatives respectively.

Our system achieves a best average F-score of 78.4 after running 10-fold cross validation using the random forest classifier on the dataset.

Features	RBF Kernel SVM
Character n-grams	73.1
Word n-grams	71.4
Sarcasm indicative tokens	66.1
Emoticons	62.8
All features	76.5

Table 6.4 F-scores for RBF Kernel SVM Classifier for Sarcasm Detection

Table 6.4, 6.5 and 6.6 shows the F-scores achieved by each of the techniques for each feature separately as well as with all the features combined.

Features	Random Forest
Character n-grams	75.0
Word n-grams	76.7
Sarcasm indicative tokens	72.0
Emoticons	68.5
All features	78.4

Table 6.5 F-scores for Random Forest Classifier for Sarcasm Detection

Features	Linear SVM
Character n-grams	66.4
Word n-grams	68.0
Sarcasm indicative tokens	70.2
Emoticons	65.7
All features	71.7

Table 6.6 F-scores for Linear SVM Classifier for Sarcasm Detection

It can be observed that each feature affects each technique differently. Word n-grams perform best with random forest classifier whereas character n-grams with RBF kernel SVM and sarcasm indicative tokens perform best with linear svm.

Chapter 7

Conclusions

We started by working on building a supervised classification system for stance and gender detection in Spanish and Catalan tweets where the target for stance was Catalan Independence. This classification system was built using various word and character level features along with support vector machine with radial basis function for classification.

With opinion mining being used in various applications today along with the abundance of on-going research in this field, it has become one of the most important tasks on big data. So we pursued further in this field and presented the first English-Hindi code-mixed dataset collected from twitter for stance identification towards Demonetisation. We explained the methods used for annotating each tweet with stance towards the target and for annotating each token with a language tag.

With the increasing usage of sarcasm in social media texts tasks like opinion mining and sentiment analysis have become much more challenging. Hence, we presented the first English-Hindi code-mixed dataset for sarcasm detection collected from twitter. We explained the methods used for collecting and annotating these tweets at both tweet level for presence of sarcasm as well as at token level for language.

After presenting the datasets for stance and sarcasm detection in English-Hindi code-mixed tweets, we described a baseline supervised classification system for both stance and sarcasm detection that was developed using the same datasets. The system uses three different machine learning techniques along with various word and character level features. These techniques are then evaluated by running 10-fold cross-validation.

There is a lot of scope for improvement in the work presented. In addition, this work can provide some much-needed groundwork for further research in other areas. Here are a few examples of future work in this area:

- The datasets presented for stance and sarcasm detection can be expanded further. For example, more tweets can be collected with stance towards other multiple targets. More tweets with sarcasm on various other topics can be added to improve the dataset for sarcasm detection.

- Both the datasets presented can be further improved by normalizing each token which in turn enhance the performance of the classification system. These datasets can also be used for developing systems for automatic language identification in code-mixed texts.
- Similar datasets can be created with other language pairs and with multiple targets for stance.
- More number of features can be explored such as POS tags, word embeddings, etc. that can help improve the accuracy of the classification system.

Related Publications

Sahil Swami, Ankush Khandelwal, Manish Shrivastava, Syed Sarfaraz Akhtar 2017. *LTRC IIITH at IBEREVAL 2017: Stance and Gender Detection in Tweets on Catalan Independence*. Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017) co-located with 33th Conference of the Spanish Society for Natural Language Processing (SEPLN 2017).

Sahil Swami, Ankush Khandelwal, Vinay Singh, Syed Sarfaraz Akhtar, Manish Shrivastava 2018. *An English-Hindi Code-Mixed Corpus: Stance Annotation and Baseline System*. CICLing 2018.

Sahil Swami, Ankush Khandelwal, Vinay Singh, Syed Sarfaraz Akhtar, Manish Shrivastava 2018. *A Corpus of English-Hindi Code-Mixed Tweets for Sarcasm Detection*. CICLing 2018.

Bibliography

[1] D. Bamman and N. A. Smith. Contextualized sarcasm detection on twitter. In *ICWSM*, 2015.

[2] U. Barman, A. Das, J. Wagner, and J. Foster. Code mixing: A challenge for language identification in the language of social media. In *CodeSwitch@EMNLP*, 2014.

[3] S. Bharti, B. Vachha, R. Pradhan, K. Babu, and S. Jena. Sarcastic sentiment detection in tweets streamed in real time: a big data approach. *Digital Communications and Networks*, 2(3):108 – 121, 2016. Advances in Big Data.

[4] B. Billal, A. Fonseca, and F. Sadat. Named entity recognition and hashtag decomposition to improve the classification of tweets. In *NUT@COLING*, 2016.

[5] C. Bosco, V. Patti, and A. Bolioli. Developing corpora for sentiment analysis: The case of irony and senti-tut. *IEEE Intelligent Systems*, 28:55–63, 2013.

[6] M. Bouazizi and T. Ohtsuki. A pattern-based approach for sarcasm detection on twitter. *IEEE Access*, 4:5477–5488, 2016.

[7] J. D. Burger, J. C. Henderson, G. Kim, and G. Zarrella. Discriminating gender on twitter. In *EMNLP*, 2011.

[8] N. Cheng, R. Chandramouli, and K. Subbalakshmi. Author gender identification from text. *Digital Investigation*, 8(1):78 – 88, 2011.

[9] A. Das and B. Gambäck. Code-mixing in social media text: The last language identification frontier? *TAL*, 54:41–64, 2013.

[10] P. S. Deshmukh, S. B. Solanke, B. B. Vachha, R. K. Pradhan, K. G. S. Babu, S. K, A. Joshi, V. Sharma, P. Bhattacharyya, T. Ptacek, I. Habernal, A. Qadir, P. Surve, and L. C. D. Silva. Review paper: Sarcasm detection and observing user behavioral. 2017.

[11] A. Fernández, L. N. Chiroque, P. Morere, and A. Santos. Sentiment analysis and topic detection of spanish tweets: A comparative study of of nlp techniques. *Procesamiento del Lenguaje Natural*, 50:45–52, 2013.

[12] C. Fink, J. Kopecky, and M. Morawski. Inferring gender from the content of tweets: A region specific example. In *ICWSM*, 2012.

[13] J. L. Fleiss and J. Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33(3):613–619, 1973.

[14] E. Forslid and N. Wike;n. Automatic irony- and sarcasm detection in social media, 2015.

[15] I. Guellil and K. Boukhalfa. Social big data mining: A survey focused on opinion mining and sentiments analysis. *2015 12th International Symposium on Programming and Systems (ISPS)*, pages 1–10, 2015.

[16] S. Gupta, P. Bansal, and R. Mamidi. Resource creation for hindi-english code mixed social media text, 07 2016.

[17] M. Hellinger and A. Pauwels. Handbook of language and communication: Diversity and change, 2007.

[18] Ho and W. y. J. Code-mixing: Linguistic form and socio-cultural meaning. *International Journal of Language, Society and Culture*, http://www.educ.utas.edu.au/users/tle/JOURNAL/issues/2007/21-2.pdf, 2007.

[19] A. Joshi, P. Bhattacharyya, and M. J. Carman. Automatic sarcasm detection: A survey. *ACM Comput. Surv.*, 50:73:1–73:22, 2017.

[20] P. Krejzl, B. Hourová, and J. Steinberger. Stance detection in online discussions. *CoRR*, abs/1701.00504, 2017.

[21] P. Krejzl, B. Hourová, and J. Steinberger. Stance detection in online discussions. *CoRR*, abs/1701.00504, 2017.

[22] C. Liebrecht, F. Kunneman, and A. van den Bosch. The perfect solution for detecting sarcasm in tweets not. In *WASSA@NAACL-HLT*, 2013.

[23] C. Liu, W. Li, B. Demarest, Y. Chen, S. Couture, D. Dakota, N. Haduong, N. Kaufman, A. Lamont, M. Pancholi, K. Steimel, and S. Kübler. Iucl at semeval-2016 task 6: An ensemble model for stance detection in twitter. In *SemEval@NAACL-HLT*, 2016.

[24] T. M., M. M.A., R. F., R. P., B. C., and P. V. Overview of the task of stance and gender detection in tweets on catalan independence at ibereval 2017. *Notebook Papers of 2nd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL), Murcia, Spain, September 19, CEUR Workshop Proceedings*, 2017.

[25] T. Mandl, M. Shramko, O. Tartakovski, and C. Womser-Hacker. Language identification in multi-lingual web-documents. In C. Kop, G. Fliedl, H. C. Mayr, and E. Métais, editors, *Natural Language Processing and Information Systems*, pages 153–163, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[26] J. Marquardt, G. Farnadi, G. Vasudevan, M.-F. Moens, S. Davalos, A. Teredesai, and M. D. Cock. Age and gender identification in social media. In *CLEF*, 2014.

[27] Z. Miller, B. Dickinson, and W. Hu. Gender prediction on twitter using stream algorithms with n-gram character features. *International Journal of Intelligence Science*, Vol. 2 No. 4A, 2012.

[28] S. M. Mohammad, P. Sobhani, and S. Kiritchenko. Stance and sentiment in tweets. *CoRR*, abs/1605.01655, 2016.

[29] C. Peersman, W. Daelemans, and L. Van Vaerenbergh. Predicting age and gender in online social networks. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*, SMUC '11, pages 37–44, New York, NY, USA, 2011. ACM.

[30] T. Ptácek, I. Habernal, and J. Y. Hong. Sarcasm detection on czech and english twitter. In *COLING*, 2014.

[31] S. Swami, A. Khandelwal, M. Shrivastava, and S. S. Akhtar. Ltrc iiith at ibereval 2017: Stance and gender detection in tweets on catalan independence. 2017.

[32] Y. Vyas, S. Gella, J. Sharma, K. Bali, and M. Choudhury. Pos tagging of english-hindi code-mixed social media content. In *EMNLP*, 2014.