AN ANALYSIS OF NEWS ON MICROBLOGGING SYSTEMS

by

Okay Aslan

B.S, Computer Engineering, Yeditepe University, 2006

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering

Boğaziçi University

2010

AN ANALYSIS OF NEWS ON MICROBLOGGING SYSTEMS

APPROVED BY:

Dr. Suzan Üsküdarlı (Thesis Supervisor)	.....
Assoc. Prof. Haluk Bingöl	.....
Assoc. Prof. Yağmur Denizhan	.....

DATE OF APPROVAL: 17.09.2010

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my supervisor Dr. Suzan Üsküdarlı for her guidance, encouragement and endless support throughout this thesis. I really appreciate all you have done for this project, without your patience and understanding; this thesis would not have been accomplished.

I am grateful to my thesis jury members Assist. Prof. Haluk Bingöl and Assoc. Prof. Yağmur Denizhan for their understanding throughout this thesis and valuable critiques that have greatly improved this project.

I am indebted to my colleague and my team leader Burcu Altın for her great understanding all through this thesis and thankful to all members of TG2304 department for their great support.

I would like to thank, all members of SosLab for their contributions and supports.

Special thanks to my housemates Onur Dinçer and Güneş Ergun for their patience and supports.

I am grateful to TÜBİTAK for supporting me with National scholarship Program for M.Sc. Students - 2210. This work also is partially funded by B.U. Research Funds (BAP 08A103 and BAP 09HA102P).

Last but the best, my special gratitude goes to my family, İsmet Aslan, Rebiye Aslan and my brother Eren Aslan for their unconditional love and endless support through my life.

ABSTRACT

AN ANALYSIS OF NEWS ON MICROBLOGGING SYSTEMS

Recent advancements in Web have enabled the wide scale participation in content, which has changed the way we communicate and access information. Traditionally people were subject to accessing news from main stream media such as newspapers and televisions. With the advent of participatory web, citizen journalism has emerged, which manifests itself on the Web. The most recent form of Web publication are microblogs. Microblogs are similar to blogs, which are timestamped posts that are most typically consumed through subscription. Unlike conventional blogs, however, microblogs differ in their tiny size (140 characters) and the frequency of posting.

Microblogs are used for a variety of reasons. Social interaction, information gathering, information sharing, marketing, and spam are among the key uses. Thought and feelings regarding global events are sure to reflect on microblogs.

The massive quantity of posts make these platforms interesting to explore with respect to news information sharing behavior. At the same time, the massive quantity also makes it a challenge to identify posts that are informative.

This thesis proposes an approach for identifying news posts. This approach is implemented in order to fetch and analyze such tweets to study temporal and quantitative properties of such posts.

- How do main stream media use microblogs?
- Do individuals share news If so, how?
- What are the temporal and quantitative properties?

Twitter, the most popular microblogging system (at the time of the writing of this thesis) was used as microblogging system. The posts (called tweets) were filtered to identify news tweets, both for news events as well as individuals. The results based on 60 users and various news events are presented.

In this thesis, a news pattern to identify news contributions was introduced. Twitter was chosen as a microblog source.A number of tweets related to different global events and individual microblog user posts were examined.

ÖZET

MICROBLOG SİSTEMLERİNDE HABER ANALİZİ

İnternet ağında yaşanan son gelişmeler, içerik bakımından geniş çaplı bir katılım sağlamaya zemin hazırlamış ve bizim iletişim biçimimizi ve bilgiye erişim şeklimizi değiştirmiştir.Önceleri habere ulaşmak için insanlar gazete ve televizyona bağlıydılar. Katılımıcı ağın ortaya çıkmasıyla, vatandaş gazeteciliği kendini İnternette göstermiştir. İnternet ağında yayın yapmanın en son örneği microbloglardır. Microbloglar, bloglara benzerler ve zaman değeri bulunan gönderilerden oluşup genellikle üyelik yoluyla kullanılırlar. Fakat, ufak boyutlarıyla(140 karakter) ve yüksek mesaj gönderme sıklıklarıyla geleneksel bloglarlardan farklılık gösteriler.

Microblogların çeşitli kullanılma sebepleri vardır. Sosyal etkileşim, bilgi toplama, bilgi paylaşımı, pazarlama ve spam önemli kullanımlar arasındadır. Küresel olaylar ile ilgili fikirler ve duygular microbloglarda yansıtılmaktadır.

Oldukça yüksek seviyedeki mesaj miktarı bu platformları haber bilgi paylaşım özellikleri açısından incelemek için ilgi çekici hale getirmektedir. Aynı zamanda, yüksek miktardaki veri, bilgi içeriği olan mesajları tanımlama açısından epey zorlayıcı olmaktadır.

Bu tez haber mesajlarını tanımlamak için bir yaklaşım önermektedir. Bu yaklaşım benzer mesajları yakalamak ve onların zamansal ve nicel özelliklerini analiz edip üzerinde çalışmak için uygulanmıştır.

- Haber sağlayıcılar mikroblogları nasıl kullanıyor?
- Bireysel kullanıcılar haber yayıyor mu ? Yayıyorsa nasıl yayıyorlar?
- Zamana ait ve niceliksel özellikler nelerdir?

Twitter, en popüler mikroblog sistemi (bu tez yazılırken) mikroblog sistemi olarak kullanılmıştır. Hem haber olayları hem de bireyler açısından, iletiler (tweets olarak adlandırılır) haber tweetlerini ayırt etmek için filtrelenmiştir. 60 kullanıcı ve çeşitli haber olaylarına dayanan sonuçlar sunulmuştur.

Bu tez çalışmasında, haber paylaşımlarını belirlemek için bir haber modeli tanıtılmıştır. Twitter microlog kaynağı olarak seçilmiştir. Bir kısım, farklı küresel olaylara ait ve bireysel microblog kullanıcı mesajları incelenmiştir.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS . . . iii
ABSTRACT . . . iv
ÖZET . . . vi
LIST OF FIGURES . . . x
LIST OF TABLES . . . xiv
LIST OF SYMBOLS/ABBREVIATIONS . . . xx
1. INTRODUCTION . . . 1
1.1. Problem Statement . . . 3
2. BACKGROUND . . . 5
2.1. Evolution Of Web . . . 5
2.2. Microblogging . . . 7
2.2.1. Twitter . . . 7
2.3. Parts Of Speech . . . 9
2.3.1. WordNet . . . 10
3. LITERATURE REVIEW . . . 11
4. APPROACH . . . 14
4.1. Preliminary Analysis . . . 17
4.1.1. Matching posts and N-V-N Pattern . . . 18
4.2. News Pattern . . . 21
4.2.1. Definition of N-V-N pattern . . . 21
4.3. Gathering and Analyzing Microblog Posts . . . 22
4.3.1. Querying Microblogging Environment . . . 22
4.3.2. Retrieving Microblog Contribution . . . 22
4.3.3. Analyzing Microblog Contributions . . . 22
4.3.3.1. Token Analysis . . . 23
4.3.3.2. Pattern Analysis . . . 25
5. IMPLEMENTATION . . . 26
5.1. Twiggle . . . 27
5.1.1. Server Side Implementation . . . 27

5.1.1.1. Querying Twitter 28
5.1.1.2. Storing Tweets in the Database 29
5.1.1.3. Analyzing tweets 32
5.1.2. Client Side Implementation 38
6. EXPERIMENTS AND RESULTS 40
6.1. Case Study: User Analysis 40
6.2. Case Study: World Cup Tweets 46
6.2.1. Argentina vs Germany 46
6.2.2. Brazil 53
6.2.3. Netherlands vs Spain 54
6.3. Case Study: Miscellaneous News 56
6.3.1. The Attack of Israel to Flotilla 56
6.3.2. Moscow Metro Bombing 58
6.3.3. Oil spill disaster in Gulf of Mexico 59
6.3.4. Other Events 60
7. CONCLUSIONS 62
7.1. Future Work and Discussions 63
APPENDIX A: STOP WORDS LIST 64
APPENDIX B: TWITTER USER DATA 66
APPENDIX C: TWITTER KEYWORD SEARCH DATA 126
REFERENCES 137

LIST OF FIGURES

Figure 2.1. Evolution Of Web 5

Figure 2.2. Examples of Web 2.0 Web Sites 6

Figure 4.1. Microblog news extraction and their properties 14

Figure 4.2. Algorithm applied to News headlines 17

Figure 4.3. Headline example which match N-V-N pattern 1 19

Figure 4.4. Headline example which match N-V-N pattern 2 19

Figure 4.5. Headline example which match N-V-N pattern 3 20

Figure 4.6. Headline example which does not match N-V-N pattern 4 20

Figure 4.7. Headline example which does not match N-V-N pattern 5 20

Figure 4.8. Contribution Analysis Steps 23

Figure 5.1. Architectural Overview of Twiggle 26

Figure 5.2. Initializing Twitter Object 28

Figure 5.3. Querying Twitter 29

Figure 5.4. The attributes of a Tweet 30

Figure 5.5. Querying User Tweets 30

Figure 5.6. Create Table SQL Query 31

Figure 5.7. Insert SQL Query 31

Figure 5.8. Select SQL Query 31

Figure 5.9. Algorithm for Analyzing Tweets 32

Figure 5.10. The attributes of a TweetAnalysisData 33

Figure 5.11. The attributes of a NvNTriplet Object 33

Figure 5.12. The algorithm for tweet pre-proccesing 35

Figure 5.13. The algorithm for getPos() function 36

Figure 5.14. The attributes of a POSData object 36

Figure 5.15. The pseudocode for finding N-V-N sequence 37

Figure 5.16. Twiggle Logo and Search Interface 38

Figure 5.17. TwiggleResult Interface 39

Figure 6.1. The distributions of all tweets and news tweets by days for keyword Argentina 47

Figure 6.2. The distributions of all tweets and news tweets by days for keyword Messi 47

Figure 6.3. The distributions of all tweets and news tweets by days for keyword Germany 48

Figure 6.4. The occurrence frequency of all nNouns in all tweets for keyword Germany 50

Figure 6.5. The occurrence frequency of all nNouns in news tweets for keyword Germany 50

Figure 6.6. The occurrence frequency of all nNouns in all tweets for keyword Argentina 51

Figure 6.7. The occurrence frequency of all nNouns in news tweets for keyword Argentina 51

Figure 6.8. The occurrence frequency of all nNouns in all tweets for keyword Messi 52

Figure 6.9. The occurrence frequency of all nNouns in news tweets for keyword Messi 52

Figure 6.10. The distributions of all tweets and news tweets by days for keyword Brazil 53

Figure 6.11. The distributions of all tweets and news tweets by days for keyword Netherlands 55

Figure 6.12. The distributions of all tweets and news tweets by days for keyword Spain 55

Figure 6.13. The distributions of all tweets and news tweets by days for keyword Flotilla 56

Figure 6.14. The distributions of all tweets and news tweets by days for keyword Gaza 57

Figure 6.15. The distributions of all tweets and news tweets by days for keyword Israel 57

Figure 6.16. The distributions of all tweets and news tweets by days for keyword MoscowMetro 59

Figure 6.17. The distributions of all tweets and news tweets by days for keyword OilSpill 60

LIST OF TABLES

Table 2.1. Functions used in Token Analysis 10

Table 4.1. Main data types of microblog systems 15

Table 4.2. Data types used in the model 16

Table 4.3. Functions used in the model 16

Table 5.1. Tools used in the implementation of Twiggle 27

Table 6.1. The categories of the user in terms of their interests 41

6.2 News tweet Ratio for selected users 41

Table 6.3. The news tweets per tweets ratio for keywords Argentina, Messi and Germany 49

Table 6.4. The news tweets per tweets ratio for keyword Brazil 54

Table 6.5. The news tweets per tweets ratio for keywords Flotilla, Gaza and Israel 58

Table B.1. The summary of the collected data for user MariahCarey 66

Table B.2. The summary of the collected data for user NBCNews 67

Table B.3. The summary of the collected data for user hrheingold 68

Table B.4. The summary of the collected data for user Mediabistro 69

Table B.5. The summary of the collected data for user mashable 70

Table B.6. The summary of the collected data for user shoemoney 71

Table B.7. The summary of the collected data for user Jason 72

Table B.8. The summary of the collected data for user Oprah 73

Table B.9. The summary of the collected data for user TheEllenShow 74

Table B.10. The summary of the collected data for user GuyKawasaki 75

Table B.11. The summary of the collected data for user journalismnews 76

Table B.12. The summary of the collected data for user bbcnews 77

Table B.13. The summary of the collected data for user snoopdogg 78

Table B.14. The summary of the collected data for user BarackObama 79

Table B.15. The summary of the collected data for user kevinrose 80

Table B.16. The summary of the collected data for user TIME 81

Table B.17. The summary of the collected data for user aplusk 82

Table B.18. The summary of the collected data for user andersoncooper 83

Table B.19. The summary of the collected data for user 50cent 84

Table B.20. The summary of the collected data for user BreakingNews 85

Table B.21. The summary of the collected data for user TheOnion 86

Table B.22. The summary of the collected data for user cnnbrk 87

Table B.23. The summary of the collected data for user TechCrunch 88

Table B.24. The summary of the collected data for user dsearls 89

Table B.25. The summary of the collected data for user davewiner 90

Table B.26. The summary of the collected data for user zappos 91

Table B.27. The summary of the collected data for user postsecret 92

Table B.28. The summary of the collected data for user TheEconomist 93

Table B.29. The summary of the collected data for user google 94

Table B.30. The summary of the collected data for user zef 95

Table B.31. The summary of the collected data for user nytimes 96

Table B.32. The summary of the collected data for user britneyspears 97

Table B.33. The summary of the collected data for user jayrosen_nyu 98

Table B.34. The summary of the collected data for user ijustine 99

Table B.35. The summary of the collected data for user PerezHilton 100

Table B.36. The summary of the collected data for user BryanAlexander 101

Table B.37. The summary of the collected data for user Scobleizer 102

Table B.38. The summary of the collected data for user smashingmag 103

Table B.39. The summary of the collected data for user THE_REAL_SHAQ . . 104

Table B.40. The summary of the collected data for user NiemanLab 105

Table B.41. The summary of the collected data for user uskudarli 106

Table B.42. The summary of the collected data for user leolaporte 107

Table B.43. The summary of the collected data for user johnbreslin 108

Table B.44. The summary of the collected data for user Veronica 109

Table B.45. The summary of the collected data for user EelcoVisser 110

Table B.46. The summary of the collected data for user iamdiddy 111

Table B.47. The summary of the collected data for user Newsweek 112

Table B.48. The summary of the collected data for user CBSNews 113

Table B.49. The summary of the collected data for user Poynter 114

Table B.50. The summary of the collected data for user TeamMalachiae 115

Table B.51. The summary of the collected data for user mathrabbit1 116

Table B.52. The summary of the collected data for user greatspeaking 117

Table B.53. The summary of the collected data for user stjohnk5 118

Table B.54. The summary of the collected data for user orangeunicorns 119

Table B.55. The summary of the collected data for user anna0974 120

Table B.56. The summary of the collected data for user iTauqeer 121

Table B.57. The summary of the collected data for user LUKIKA 122

Table B.58. The summary of the collected data for user Alyssafelldown 123

Table B.59. The summary of the collected data for user Hajji_love 124

Table B.60. The summary of the collected data for user CallaLove 125

Table C.1. The summary of the collected data for keyword Moscow Metro . . 126

Table C.2. The summary of the collected data for keyword spain 127

Table C.3. The summary of the collected data for keyword germany 128

Table C.4. The summary of the collected data for keyword oilspill 129

Table C.5. The summary of the collected data for keyword messi 130

Table C.6. The summary of the collected data for keyword israel 131

Table C.7. The summary of the collected data for keyword argentina 132

Table C.8. The summary of the collected data for keyword flotilla 133

Table C.9. The summary of the collected data for keyword netherlands 134

Table C.10. The summary of the collected data for keyword brazil 135

Table C.11. The summary of the collected data for keyword gaza 136

LIST OF SYMBOLS/ABBREVIATIONS

*	The regex * (zero or more occurrences)
+	The regex + (one or more occurrences)
nNoun	Noun or Named Entity
N-V-N	nNoun-Verb-nNoun
POS	Part of Speech

1. INTRODUCTION

Collaborative web platforms where people interact with each other are one of the key milestones of the evolution of the internet. The human factor plays a significant role in the development of such platforms by providing content to the platform environment itself. Especially the web platforms where people share their opinions, ideas such as social networks, microblogs and wikis are having millions of contributions in a day.

Microbloging is one of the most remarkable recent internet trend and gaining increasing popularity among netizens. A microblog is a type of blog which permits its users to publish limited length of text. Typically a microblog environment has following important aspects:

- The limitation in the size of a microblog contribution provides a suitable environment for users to make their submissions via various kinds of clients like mobile phones and web applications. Thus, it is easy to submit contributions to microblogs.
- The simplicity of accessing and updating a microblog leads to a frequent use of microblog services. Increasing number of microbloggers contribute to microblogging systems either reading microblog contributions or submitting microblog posts. This results in millions of contributions per day and makes the microblogging systems a rapid information sharing platform.
- Microblogs are generally open platforms and a microbloger can submit content free posts to microblog spaces. Hence, a microblog contribution may contain different kinds of posts in terms of their contents. Some microblog posts convey what the microblogger is doing or thinking, some express news from an event or comment to an existing event and some commercial posts aim to advertize services or products or to sponsor collaboration within an organization.

As a result, in the case of spreading information, truth or fiction microblogs are probably the most famous ones nowadays.

Since microblogs are fast, easy and open, they can be thought as a news distribution platform for both news originator and news reader. When talking about news on microblogs, mainly, we can mention two types of news contributions in terms of their source. First one is the mainstream media news which is disseminated via the largest distribution channels like newspapers or televisions. The originators of this type of news are generally microblogger bots which post news from mass media to microblog environment. The second one is the non-mainstream media news which is sourced in and spread through microblogs. The originators of this type of news are human microbloggers and these types of news are local ones which are mostly not on the public eye but may interest other microbloggers.

In this study, a model which extracts news contributions from microblogs is proposed and some news analyses are performed in the data gathered from the following categories.

- Individual microblog users post : The contributions of some selected microblog users were analyzed and the users were ranked according to their news contribution ratios.
- Global Events : Some selected global events were followed and some quantitative news analysis was performed on numerous microblog contributions related to selected events.

The main goal of this thesis is to understand news sharing characteristics of microblogs by comparing microblogs with other news platforms such as mass media.

In the following sections, first of all, brief information about background concepts and related works are explained. Then, the proposed model and the implementation of it are described. At the end, experiments and results and conclusions are mentioned.

1.1. Problem Statement

Microblogs are the blogs those allows users to send brief text updates. (eg: Twitters allows 140 characters or fewer). The text size limitation of a microblog post makes the essential difference between microblogs and traditional blogs. Millions of posts have been submitted everyday to microblogging environments.

People utilize microblogs for both business and individual reasons. A microblogger can post anything freely to a microblog. Many microblog posts are created on person-to-person level to chitchat or to spam to promote a company's products or to share news about global or local events.

Nowadays, mobile and easy to use microblogs, provides the most preferred environment for news distribution. A news in a microblog post can be a mainstream media news like the news issued via newspapers or televisions which may interest people all over the world or can be a non-mainstream news such as news about a local event which only interest native people who live where the event takes place.

The following microblog post which is about the Gulf of Mexico oil spill disaster can be count as an example of main-stream media news.

> RT @AJEnglish: BP begins capturing Gulf oil spill: Energy giant begins funnelling oil onto ship after cap is placed on leaking pipe. http://aje.me/cFoCrm

On the other hand the following is a non-mainstream media news post which mentions a workshop in a web privacy conference.

> Kasey Chappel at W3C privacy workshop; social, technical and economic changes driving issue related to privacy #w3privacy

The aim of this thesis is to make analysis on the news posts in the microblogs to gain insight about news sharing characteristics of a microblog by answering following questions:

- How do main stream media use them?
- Do individuals share news ? If so, how?
- What are the temporal and quantitative properties?

Since not all microblog posts contain news information, one of the most crucial problem encountered in this thesis is to distinguish news related posts from other posts. To address this issue, a model to extract posts which are most likely to be a news post is proposed.

2. BACKGROUND

In the section, some important background concepts and technologies are described to provide a better understanding of the remainder of the thesis.

2.1. Evolution Of Web

The following Figure 2.1 simply describes the evolution of web.

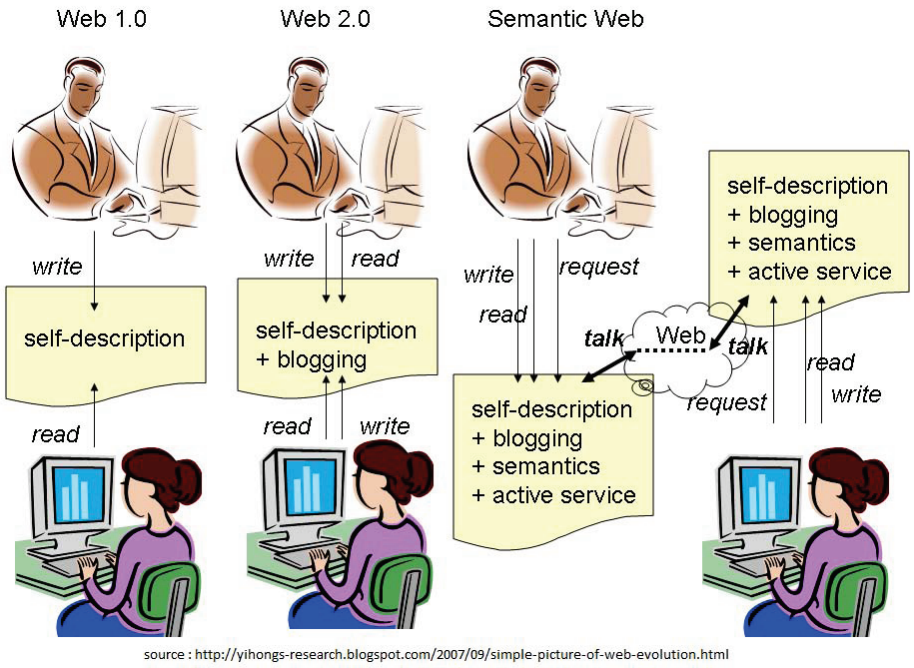


Figure 2.1. Evolution Of Web

The first stage of the World Wide Web is the Web 1.0. It is the preliminary era of the World Wide Web and particularly consists of static web pages where authors of web pages publishes what they want to share and the readers can visit these web pages and read the shared info. There is no direct connection between web site authors and readers. Shortly, in Web 1.0, the authors just write and the readers just read.

After Web 1.0, the World Wide Web was evolved to Web 2.0. In Web 2.0, both

web page authors and the readers can share the same web space and read/write on this space. This is the most important difference between Web 1.0 and Web 2.0 and provides a suitable environment to establish friendly social communication among web users. This advancement increases the participating interest of web users and the web users, most of the time, become connected with each other in Web 2.0. Having millions of subscribers the social friendship networks like Facebook [1] and MySpace [2], collaborative information sharing web sites like Wikipedia [3] and microblogging web sites like Twitter [4] are the most crucial examples of the Web 2.0. Figure 2.2 shows some example web sites of the Web 2.0.



Figure 2.2. Examples of Web 2.0 Web Sites

The next stage of the World Wide Web is believed as Semantic Web, where the idea is to create machines to understand and process the data available on the Web. These machines can execute owner-specified requests by themselves and communicate with other semantic web spaces. The early Semantic Web efforts consist of languages specifically designed for defining and querying data, such as eXtensible Markup Language (XML) [5], Resource Description Framework (RDF) [6], Sparql Protocol And Resource Query Language (SPARQL) [7].

2.2. Microblogging

Microblogging is one of the most prominent outcomes of the Web 2.0. It is a type of blogging mechanism where microblog users contribute to the microblog environment by short text updates. The basic difference between traditional blogging and microblogging is the text size limitation of the microblog post. A microblogger is forced to submit its contribution via a limited size of text message.This makes microblogs updatable via sending SMS from mobile clients like cell phones. Since microblogs can easily be accessible and updatable via mobile clients, microblogging becomes fairly popular in todays internet and its use is growing.

As microblogs consist of content-free posts, microbloggers use microblogs for different purposes. Some users express their daily activities; some delivers their ideas on something, and some broadcast news or promote something whereas some prefers to chat with other microblog users. Increasing popularity of microblogs attracts celebrities, politicians and companies as well. They generally use microblogs to reach large audience.

Microblogs are publicly open, rapid and easy to use, and the update rates of the microblogs are tremendous if we think of millions of subscribers. Eventually, these contributions forms a valuable and unprocessed microblog post cloud and this massive data interests academicians and become subject to many studies.

There are many microblogging services in the web like Twitter [4], Tumblr [8], Jaiku [9]. The most popular of them is Twitter [4].

2.2.1. Twitter

Twitter which is the most well-known microblog, was created in 2006. Its popularity topped after March 2007 when it achieves the Web Awards of South by Southwest [10] conference. After this award, number of microblog posts send to twitter increased from 20000 to 60000 per day.

According to data announced at the official Twitter Developer Conference - Chirp which was take place at San Fransisco on April 2010 [11], Twitter has the following statistics:

- Number of registered users is 105,779,710
- Number of new users sign up per day is 300,000
- Approximately 60 per cent of new users are coming from outside the U.S
- Number of unique visitors per month is 180 million
- 75 per cent of Twitter traffic comes from third-party applications
- 60 per cent of all tweets sent to Twitter come from third-party apps
- Number of search queries on Twitter per day is 600 million
- Twitter has 175 employees
- Number of Twitter applications is over 100,000
- Number of requests a day through Twitter API 3 billion
- 37 per cent of active Twitter users use their phone to tweet

The microblog posts send to the Twitter is called as "tweet". A tweet consists of plain text which has 140 character limitation. The followings are the special keywords used in a tweet which has different meanings:

- @*username* : This type of keywords are called mentions. If a twitter user wants reference or reply to another user, uses this pattern. (eg: @Bob, @David)
- #*tag* : This type of keywords are called hashtags. Twitter users uses this pattern to tag a tweet. Hashtags can be found anywhere in a tweet.They are used to show an indication of what the tweet is about and put by the creator of the tweet. (eg: #apple, #google)
- RT : This keyword is used to repeat a tweet. A RT sign in the beginning of a tweet shows that the tweet is a re-post of another tweet. A Twitter user retweets a tweet when they share the idea of the tweet or wants to spread it.(eg: RT What a lovely day!!!)

A twitter user can follow another twitter user without any permission from the

followed user. In this case the updates submitted by the followed user are shown in the follower user's home page. The number of followers rate of the celebrities or the politicians can be quite high.(eg: Barrack Obama has 4.8 million, Ashton Kutcher has 5.4 million followers in August 2010)

Some reputable mass media like CNN [12], BBC [13], New York Times [14] etc. also publish news automatically on the twitter.

To allow Twitter to be accessible and updatable via 3rd party mobile or web clients twitter supports an API, so that using this API [15], one can post tweets, receive update, follow other users etc. One of the most popular Twitter API is twitter4j [16] which is also used during implementation of proposed model.

The followings are some random chosen example tweets from Twitter.

- Wondering how I can comprehend if using NAS as a file server is preferable to a dedicated computer. Thinking of 4 TB RAID
- Inception tonight, high expectations!
- Tiger Woods finishes worst tournament as professional at 18 over par. http://on.cnn.com/ctOfva
- RT An#annoyingquestion is 'what's wrong?' when you're completely fine.

2.3. Parts Of Speech

In English Grammar, the words are classified into eight parts of speech(POS): verb, noun, pronoun, adjective, adverb, preposition, conjunction, and interjection. [17]

The part of speech of a word explains how the word is used in the sentence. A word can be noun in one sentence but verb in another.

The followings are example sentences:

A **CPU** is made of metal and silicon.

Here CPU is noun.

They **walk** down the street.

Here walks is verb.

The hospital stands on the **walk** to the house.

Here walks is noun.

2.3.1. WordNet

WordNet [18] is a lexical database of English language. George A. Miller who is a professor of psychology at Princeton University's Department of Psychology is the principal investigator of WordNet. In WordNet, such parts of speech of the words as nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets). WordNet provides short and general definitions of these synonym sets. WordNet can be utilized to support automatic text analysis and artificial intelligence applications. It is also freely available for download. In this thesis, WordNet is utilized to determine parts of speech of a given word. The followings are some example words and their part-of-speechs from WordNet.

Table 2.1. Functions used in Token Analysis

Word	**POS**
book	Noun,Verb
great	Noun,Adjective
receive	Verb
apple	Noun

3. LITERATURE REVIEW

As the popularity of Microblogs grows among netizens, the interests of academicians on microblogs also grow correspondingly. Enormous number of microblog posts provide different research areas for researchers. Since microblogging is a new concept, the studies on microblogs are new as well.

Twitter, which is the most famous microblog attracts various community of users with different interests. The work in [19], observes the topological and geographical properties of Twitter's social network and analyzes user intentions and shows how users with similar intentions connect with each other. They made informative observations like:

- Most of the tweets in Twitter talk about daily routine or what people are currently doing.
- About one eighth of all tweets in their collections contain mentions.
- About 13 per cent of all tweets in their collection contain some URL in them.
- Most Twitter users report latest news or comment about trendy events. Some automated Twitter users' tweets like weather reports and news stories from RSS feeds
- Twitter users can be categorized as,
  (i) automated or human agents, who tweets frequently valuable and tend to have a large number of followers,
  (ii) users who just follow friends, family and co-workers.
  (iii) users who tweet rarely, but follow other users regularly.

Balachander Krishnamurthy et al. [20], analyze twitter dataset covering nearly 100,000 users and categorise as broadcasters who tend to tweet a lot including news sources such as New York Times and BBC, Acquaintances whose following/followed relationships close each other and Miscreants including spammers who contact everyone they can hoping that some will follow them.

Twitter is great tool for distributing information, the examples [21] and [22] show how people use twitter and express their opinions on the events Southern California earthquake and 2009 Iraninan election respectively.

In [23], Sarah Vieweg et al. study communications published by people during natural disasters. They analyze microblog data during Oklahoma Grassfires on April 2009 and the Red River Floods in March and April 2009. They suggest that the features of the information gathered during emergency situations leads the development of software systems that use information extraction strategies.

Shamma et al. [24] focus on the Twitter usage in live events. They analyze Twitter data during the 2008 Presidential Debates. They examined the structure of Twitter traffic, and found out that the structure of Twitter traffic can provide as a predictor of changes in topics in the media event. On the other hand they suggest that Twitter posts can reflect the topics of discussion in the mass media.

In [25], Haewoon Kwak et al. perform a comprehensive and quantitative study on twitter by collecting millions of user profiles, social relations and tweets to discuss the characteristics of twitter as a social network and news Media. They studied on follower-following topology of Twitter users to analyze characteristics of human social networks. They also classified the trending topics based on the active period and show that the majority of topics are headline news or persistent news in nature.

In a study, Owen Phelan et al. [26] propose an approach to news recommendation utilizing real-time microblogging activity. They use twitter as microblog service and promote news stories from a user's favorite RSS feeds. In this model users give their favourite RSS feeds to the system and system promotes the feeds based on the co-occurence of popular terms within the user's RSS and Twitter indexes. Their proposed model is promising and shows that users can benefit from the recommendations that are derived from the Twitter data.

In [27], Jagan Sankaranarayanany et al. suggest a system called TwitterStand that captures tweets those correspond to late breaking news. They use a set of tweet seeders to feed their system and use a naive Bayes classifier to extract news from other tweets. They find out that their results are analogous to a distributed news wire services. They particularly mention Twitter as a noisy environment and it is hard to differentiate news tweets from other tweets.

The following studies are performed on Twitter, by my colleagues from Bogaziçi University:

- Ece in [28], focuses on the tags used in the Twitter network to analyze user contributions to microblogs. Particularly, she studied on word-tag, word-user and tag-user relationships and associated users who have common interests based on their tag usage. She argued that, her model can be used for user suggestion based on common interest.
- Emre in [29], categorises Twitter users according to words used in their tweets. He analyzes occurrence frequency of the words in the tweets and benefited from DBPedia as a semantic resource while categorising users. He suggested a group of category names for Twitter users which can also be used as a user suggestion tool.
- Duygu in [30], proposes a model to reveal Twitter user characteristic and interests. She, first selected 150 users from Twitter based on different categories and then compared them by studying on their tweets. She analyzed internal and external references in the tweets like web URLs, hashtags etc and frequency of occurrence of the significant words in the tweets. She, finally, tried to observe common characteristic of some selected user groups.

The above three studies commonly focus on user tweets and collect tweets from some particular users and analyze their tweets. On the other hand, in this work, both tweets from some selected users and tweets from public timeline are analyzed.

4. APPROACH

This chapter describes the proposed model in order to extract news contributions from a set of microblog contributions using a news pattern. Figure 4.1 shows the overall system design.

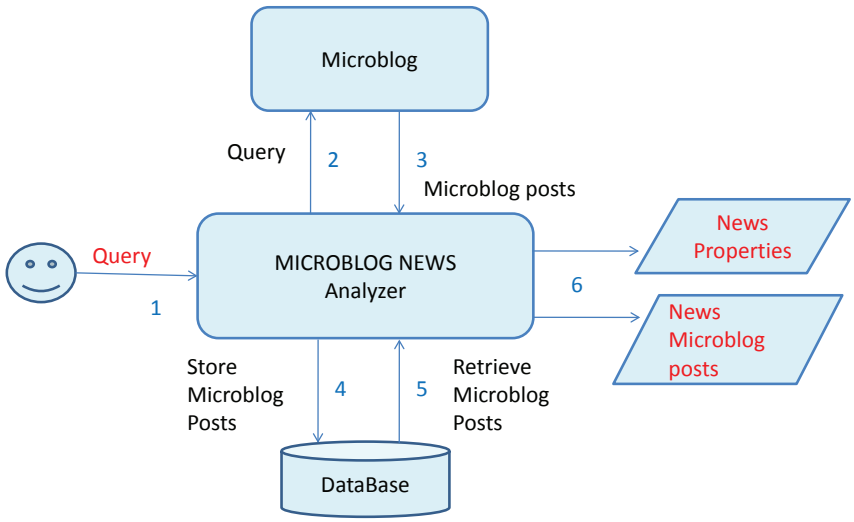


Figure 4.1. Microblog news extraction and their properties

According to this model, a user provides a key phrase query which is used to retrieve a set of microblog posts.This set of contribution is stored in the database for later use. The stored contributions are analyzed to determine if they match a "News Pattern". Here, "News Pattern" is a pattern which can be applied to microblog contribution to decide whether a contribution is likely a news contribution. All contributions that match the pattern are considered as news posts. The output is a set of news microblog posts and related characteristic.

The significant data types and functions used in the model are given in the tables 4.1, 4.2, and 4.3.

Table 4.1. Main data types of microblog systems

Type Name	**Specification**
MicroblogSystem	$\langle Microbloggers, Microblogs \rangle$
Microblogger	$\langle Name, microbloggerId \rangle$
microbloggerId	$Integer$
Microbloggers	$\{mbr_1, mbr_2, ...\} \mid mbr_i : Microblogger$
Microblog	$\langle Microblogger, Posts \rangle$
Microblogs	$\{m_1, m_2, ...\} \mid m_i : Microblog$
Contribution	is the text of the post
Contributions	$\{c_1, c_2, ...\} \mid c_i : Contribution$
Post	$\langle Microblogger, postId, Contribution, TimeStamp \rangle$
postId	$Integer$
Posts	$\{p_1, p_2, ...\} \mid p_i : Post$

Table 4.2. Data types used in the model

Type Name	**Specification**
Token	is a sequence of non space characters
Stopwords	$\{st_1, st_2, ...\} \mid st_i : Token$
Keyword	$\{k_1, k_2, ...\} \mid k_i : Token$
Noun	$n \mid getPos(n) =$ "$Noun$"
Verb	$v \mid getPos(v) =$ "$Verb$"
NamedEntity	$ne \mid getPos(ne) =$ "$Unidentified$"
PartOfSpeech	"$Noun$" $\vee$ "$Verb$" $\vee$ "$Adverb$" $\vee$ "$Adjective$" $\vee$ "$Unidentified$"
InSigWord	$s \mid s \in Stopwords \vee$ "$Adverb$" $\vee$ "$Adjective$"
InSigWords	$\{inS_1, inS_2, ...inS_n\} \mid inS_i : InSigWord$
SigWord	"$Nous$" $\vee$ "$Verb$" $\vee$ "$Unidentified$"
SigWords	$\{s_1, s_2, ...sn\} \mid s_i : SigWord$
nNoun	$\{n_1, n_2, ...n_n\} \mid n_i : Noun \vee n_i : NamedEntity$
NewsPattern	is a pattern consisting of sequence of part-of-speech of tokens
NewsPost	$c \mid c : Contribution \wedge isNews(c, NP) = true, NP : NewsPattern$
NewsPosts	$\{newsp_1, newsp_2, ...newsp_n\} \mid newsp_i : NewsPost$

Table 4.3. Functions used in the model

Function Name	**Output**
$isNews(s : SigWords, NP : NewsPattern)$	boolean
returns true if s matches pattern NP (see Section 4.2)	
$getPos(t : Token)$	PartOfSpeech
returns the "part-of-speech" of the t (see Section 2.3)	
$getNewsPosts(k : Keyword, NP : NewsPattern, p : Posts)$	NewsPosts
returns the subset of p that matches pattern NP	

4.1. Preliminary Analysis

Before analyzing news from microblogs, a set of headline news items were inspected. 1,250 headline news items were retrieved from "Affective Task" [31], in which this set was used to classify emotions and valences in headline news text.

In the preliminary analysis, the algorithm in Figure 4.2 was applied all 1,250 headlines in the set.

```
for all headline in the corpus do
  tokenList = tokenize(headline);
  tokenList = removeStopwords(tokenList);
  for all token in the tokenList do
    determinePos(Token);
  end for
end for
```

Figure 4.2. Algorithm applied to News headlines

After examination of each headline, it was observed that 1,038 of them have the following pattern:

N-V-N

where,
N is a *Noun* or *Named Entity*
V is a *Verb*

According to [32] "Named Entity" values can be defined as a Person, an Organization, a Location, a Facility, or a Geo-Political Entity (GPEs). In the remainder of this thesis **nNoun** will be used to represent a *Noun* or a *Named Entity*.

An examination of the headline news tokens revealed that 80% of the corpus contain at least one N-V-N sequence. That is, a nNoun entity is followed by a verb and that verb is followed by a nNoun entity. Note that, the follow relationship is not necessarily adjacent and it is possible that multiple N-V-N sequences exist in a single headline.

4.1.1. Matching posts and N-V-N Pattern

The following matching rules are utilized while searching the news headlines having N-V-N pattern:

(i) If a text does not contain at least two nNouns and at least one verb, there is no match.

(ii) To match to an N-V-N pattern, a text must contain at least one nNoun-Verb-nNoun sequence. In an nNoun-Verb-nNoun sequence, there should be two nNouns with at least one verb between them. The nNouns and the verb are not required to be located adjacent in an nNoun-Verb-nNoun sequence:

(iii) There may be tokens that are not nNouns or verbs in amidst the N-V-N sequence. For instance, there may be an adjective between the first nNoun and the verb of an N-V-N sequence. Since only a verb between two nNouns can form an N-V-N sequence, non-nNouns or non-verbs such as adjectives and adverbs are ignored.

(iv) If there are multiple nNouns and verbs in the text, the elements of the nNoun-Verb-nNoun sequence are selected according to the following rules:

- If there are more than one verbs between the first and second nNoun, the last verb is chosen as the verb of the N-V-N sequence.
- If there are more than one nNouns before the verb of the N-V-N sequence, the last nNoun is chosen to matcth the first nNoun of the sequence.
- If there are more than one nNouns after the verb of the N-V-N sequence, the first nNoun is chosen to match the second nNoun of the sequence.

(v) Multiple N-V-N sequences are ignored since it is irrelevant to determining whether a posts matches an N-V-N pattern.

The Figures 4.3, 4.4 and 4.5 show example news headlines from [31] which match the N-V-N pattern.

Example 1:

"***Youtube*** ***yields*** *surprising* ***payback***"

N V N

where,

Youtube : Named Entity

yields : Verb

payback : Noun

Figure 4.3. Headline example which match N-V-N pattern 1

Example 2:

"***South Africa*** ***mourns*** ***Adelaide Tambo***"

N V N

where,

South Africa : Named Entity

mourns : Verb

Adelaide Tambo : Named Entity

Figure 4.4. Headline example which match N-V-N pattern 2

Example 3:

"***University*** *offers* ***support*** *to* ***New Orleans school***"

N V N

where,

University : Noun

support : Verb

New Orleans school : Named Entity

Figure 4.5. Headline example which match N-V-N pattern 3

The Figures 4.6 and 4.7 show example news headlines which do not match the N-V-N pattern. Such patterns for these news were not considered in this thesis.

Example 4:

"An Elephant Crackup ? "

Figure 4.6. Headline example which does not match N-V-N pattern 4

Example 4:

"Inconvenient truths"

Figure 4.7. Headline example which does not match N-V-N pattern 5

4.2. News Pattern

Given the headline news analysis and that microblog posts are very small, an examination of the use of the same N-V-N pattern was applied to microblog systems.

The explorations was to determine whether microbloggers share news in a similar manner. This was done by examining:

- post matching keywords known to be newsworthy
- post of main stream news media providers
- post of users from different categories

All of the above yield a set of posts. The characteristic of these posts analyzed.

4.2.1. Definition of N-V-N pattern

N-V-N pattern is a pattern used by the system to extract news contributions from a set of contributions, where N-V-N denotes a sequence of tokens of nNoun, verb, nNoun(recall section 4.1). The followings explain how a contribution matches the N-V-N pattern. Let's say a contribution C, consists of a number of tokens.

$$Token +$$

A contribution that matches the N-V-N pattern can be described with the following regular expression.

$$(Token*)\ nNoun\ (Token*)\ Verb\ (Token*)\ nNoun\ (Token*)$$

4.3. Gathering and Analyzing Microblog Posts

4.3.1. Querying Microblogging Environment

A key phrase known to be newsworthy is provided to the system. All posts which contain that key phrase are retrieved from the microblog system. For example key phrase "Barcelona" may be used to query a microblog system to retrieve news about a festival in Barcelona. The key phrase can be more than one word as well. For example key phrase "Oil Spill" can be used to retrieve news about the petroleum tanker accident in Gulf of Mexico.

4.3.2. Retrieving Microblog Contribution

According to the query size limits of the microblogging environment, a number of latest microblog posts, which include the key phrase, are retrieved. From a microblogger post, the following data is gathered.

- postId
- the text of the post.This is the actual text posted by the microblogger (Contribution)
- time stamp of the post
- microblogger name
- microbloggerId

All of the above data makes up the post information in a microblog. In this thesis, a "contribution" is the actual text posted by the microblogger. After querying the microblog a set of Contribution is retrieved.

4.3.3. Analyzing Microblog Contributions

Figure 4.8 shows the microblogger contribution analysis steps where each step is numbered in the order as they are processed.

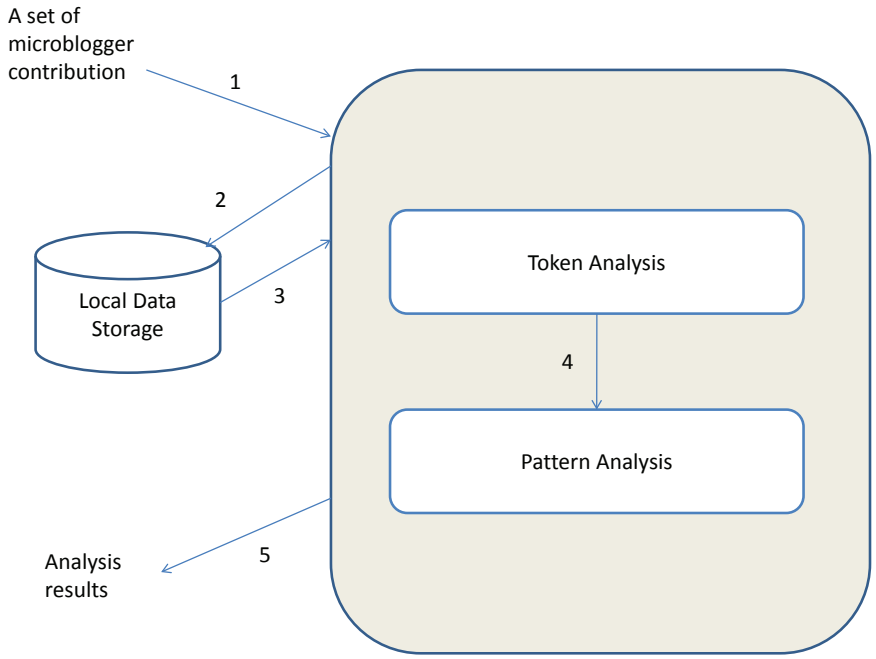


Figure 4.8. Contribution Analysis Steps

Microblog contribution analysis basically, consists of the following two processes where they are processed steps:

- Token Analysis
- Pattern Analysis

4.3.3.1. Token Analysis. The retrieved microblogger contributions processed by the token analysis process.

This step, tokenizes the posts and removes insignificant tokens from the contribution. It also computes the cumulative frequencies of the nNouns present in the posts. It counts all nNouns in all posts as well as all nNouns in all posts that match N-V-Npattern.

The insignificant tokens(InSigWords) which will not be used in the later analysis processes are the followings:

- Adverbs
- Adjectives
- Stop Words: These are the words frequently used in a natural language which do not have distinguishing property(Stopwords)

If a token is identified as noun by getPos function, it is kept as nNoun. If a token cannot be identified via getPos function, this token is taken into account as a named entity candidate and kept also as nNoun. In this model, the significant type of tokens which will be employed in the later analysis processes, are the followings:

- Noun
- Verb
- Unidentified : these are assumed as named entity candidates

In the token analysis step; system parses each contribution into tokens and for each token the following operations are performed:

- Check if the token is in Stopwords. If so, omit it.
- Using the getPos function retrieve the pos of the token.
  If it is "Adjective" or "Adverb", omit it.
  If it is Verb, keep it.
  If it is Noun or Undefined, keep it as nNoun.

After removing InSigWords from a contribution, a sequence of tokens which are considered as significant are obtained. The existence order of tokens in a contribution is preserved.

This process can be represented as the following function:

$$\text{analyzeToken}(\{c_1,c_2,\ldots,c_n\},InSigWords):=\{(c_1,(c_1 \setminus InSigWords)),$$
$$(c_2,(c_2 \setminus InSigWords)),\ldots,(c_n,(c_n \setminus InSigWords))\}$$

where,

c_i : Tokens which are tokenized from a contribution

$c_i \setminus InSigWords$: Tokens which are considered as significant

4.3.3.2. Pattern Analysis. The aim of this step is to expose contributions which match N-V-N pattern. The following operation is performed for each significant token sequence.

- Using isNews function, check if the token sequence contains the N-V-N pattern. If so, keep the original contribution and the tokens that form the NvN pattern as a pair.

This process can be represented as the following function, where the result of this function is a new set of contributions and its corresponding NvN pairs. These contributions are referred as news contributions.

$$\text{analyzePattern}(S_{set}) := \{(c_1,\ N_{11}V_1N_{12}),\ (c_2,\ N_{21}V_2N_{22}),\ \ldots,\ (c_n,\ N_{n1}V_nN_{n2})\}$$

where,

S_{set} : analyzeToken(Contributions,InsigWords)

c_i : isNews(c_i \ InSigWords, NVN)

N_{i1}, V_i, N_{i2} : Tokens $\wedge$ $pos(N_{i1})$=$pos(N_{i2})$=Noun $\wedge$ $pos(V_i)$ =Verb

5. IMPLEMENTATION

This chapter explains the implementation of the proposed model described in chapter 4. The model is designed based on traditional model-view-controller (3-tier architecture).

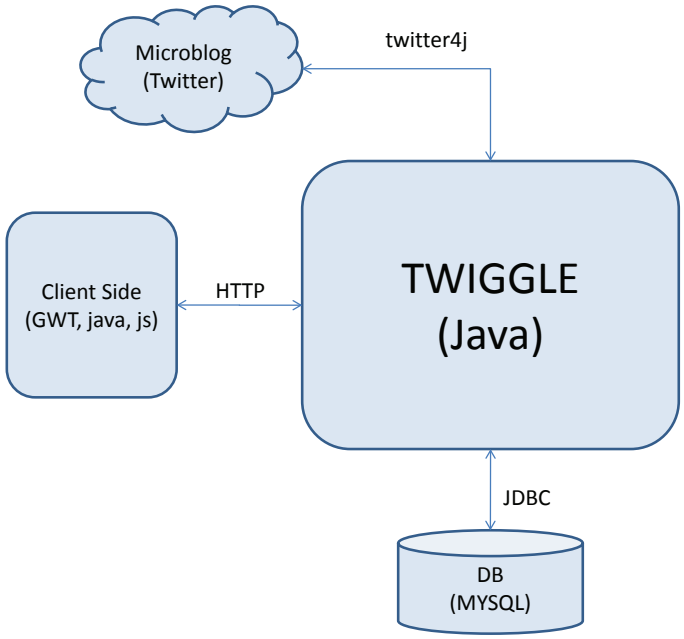


Figure 5.1. Architectural Overview of Twiggle

Figure 5.1 shows the general architectural background of the implemented system. Here the system is referred as "Twiggle". Twiggle is the main controller of the system and organizes all background operations. Twiggle is implemented by JAVA programming language. Because of its popularity and easy to use api Twitter was chosen as microblog source and Twitter4j [16] API was used to query Twitter.

Table 5.1 shows the tools used in the implementation of the system.

Table 5.1. Tools used in the implementation of Twiggle

Programming language	Java
Programming editor	Eclipse
Client Side Programming tool	Google Web Tool Kit
Database	Mysql
Part of Speech Tool	WordNet 2.1

In the following sections the implementation is explained step by step.

5.1. Twiggle

Twiggle can be thought as the heart of the implemented program. It organizes all background operations. Twiggle, in background, is a web server so its implementation can be divided into server side and client side implementations.

5.1.1. Server Side Implementation

The server side of the Twiggle is responsible for the implementation of the incoming web service request. As a web server Twiggle exposes following web service methods:

(i) searchAndReturnResults
(ii) performAnalysis

(i) The **searchAndReturnResults** method takes keyword as a parameter and basically its jobs are the followings:

- query twitter with the given keyword and retrieve a set of tweets from Twitter which contains the keyword
- storing tweets in the database

- perform analysis on each tweet to check if the tweets match N-V-N pattern
- return all tweets those match the N-V-N pattern

In the following subsections these operations are described in detail.

5.1.1.1. Querying Twitter. To query Twitter, Twitter4j [16] which is a very useful API was used. Figure 5.2 shows the java code snippets for the initialization of twitter object provided by twitter4j API. To initialize twitter object username and password of a valid Twitter subscriber account is needed.

```
Twitter twitterObj = null;
twitterObj = new TwitterFactory().getInstance("userName", "password");
int pageNum = 15;
```

Figure 5.2. Initializing Twitter Object

After initializing of twitter object, Twitter is queried. Twitter web service limits its query method to return maximum 1500 tweets per search request. Thus when Twitter is queried maximum 1500 tweets can be retrieved at a time. Figure 5.3 shows the implementation of the query Twitter operation. This implementation requires pagination and since one page is allowed maximum 100 tweets by Twitter API, the page size is chosen as 15 to reach the maximum overall limits of 1500 tweets. Since the proposed model was built based on English language, the language of the query is chosen as English, so that only the tweets written in English will be returned.

Note that number of tweets retrieved after querying Twitter is not necessarily be equal to 1500, there could be less results according to searched keyword. After querying Twitter a set of tweets containing the given keyword are achieved. Each tweet contains a set of attributes as shown in figure 5.4. Here, tweetID is the Twitter wide unique id of the tweet, userID is the Twitter wide unique id of the Twitter user who post the tweet, userName is the name of the of the Twitter user who post the tweet, source shows from where the tweet is sent (mobile, web etc.), tweetText is the actual text of the tweet,

```
List allTweets = new ArrayList();
Query tweetQuery = new Query("keyword");
tweetQuery.setLang("en");
tweetQuery.setPage(15);
tweetQuery.setRpp(100);
for i = 1 to 15 do
   tweetQuery.setPage(15 - i);
   listTweets = twitter.search(tweetQuery).getTweets();
   for j = 1 to listTweets.lenght() do
      allTweets.add(listTweets.get(j));
   end for
end for
```

Figure 5.3. Querying Twitter

tweetDate is the post date of the tweet, imageUrl is the url of the avatar used by the user who post the tweet, if the tweet is replied to another twitter user, toUserID shows the Twitter wide unique id of replied user and toUserName is the name of the replied user.When a set of tweets are collected, they are stored in the db for later use.

Implementation of querying user tweets is almost same for search query.Figure 5.5 shows the implementation of the query user tweets operation.

5.1.1.2. Storing Tweets in the Database. The retrieved tweets are stored in the database before performing analysis on them. MYSQL which is an open source and easy to use database is used for storing tweets. Whenever a new set of tweets are required to be stored in the database, a new table is created for in the name of the queried keyword.

Figure 5.6 shows the sql query used while creating a database table and Figures 5.7 and 5.8 shows the insert and select queries, respectively. When tweets are retrieved from DB, they are stored in the tweet object shown in figure 5.4 and all further observations are done on this object.

```
Tweet
(
Long tweetID = null;
int userID = 0;
String userName = null;
String source = null;
String tweetText = null;
Date tweetDate = null;
String imageUrl = null;
int toUserID = 0;
String toUserName = null;
)
```

Figure 5.4. The attributes of a Tweet

```
List allTweets = new ArrayList();
Paging paging = new Paging(32, 100);
tweetQuery.setPage(32);
tweetQuery.setRpp(100);
for i = 1 to 32 do
   paging.setPage(32-i);
   listTweets = twitter.getUserTimeline(userName, paging);
   for j = 1 to listTweets.lenght() do
      allTweets.add(listTweets.get(j));
   end for
end for
```

Figure 5.5. Querying User Tweets

```
CREATE TABLE tw_keyword_tb
(
tweet_id double NOT NULL
fromuser_id double NOT NULL
touser_id double DEFAULT NULL
tweet_text varchar(250) NOT NULL
fromuser_name varchar(100) NOT NULL
touser_name varchar(100) DEFAULT NULL
imageurl varchar(250) DEFAULT NULL
date datetime NOT NULL
source varchar(200) DEFAULT NULL
PRIMARY KEY (tweet_id)
)
```

Figure 5.6. Create Table SQL Query

```
INSERT INTO tw_keyword_tb (tweet_id, fromuser_id, touser_id,
tweet_text, fromuser_name, touser_name, imageurl, source, date)
VALUES (?,?,?,?,?,?,?,?,?);
```

Figure 5.7. Insert SQL Query

```
SELECT * FROM tw_keyword_tb;
```

Figure 5.8. Select SQL Query

```
for tweet to tweetList.length() do
    TweetAnalysisData tweetAnalysisData = preProcessTweet(tweet);
    postProcessTweet(tweetAnalysisData);
end for
```

Figure 5.9. Algorithm for Analyzing Tweets

5.1.1.3. Analyzing tweets. After a set of tweets are stored in the database, they are retrieved for analysis operation. The analysis operation can be divided into two operations as pre-processing and post-processing.

Figure 5.9 shows the algorithm implemented for the tweet analysis. According to this algorithm, each tweet in the tweetlist is first pre-processed and than post processed. The tweetlist is an array of tweets which are previously retrieved from Twitter and stored in the database. TweetAnalysisData is a class that stores various analysis data for tweet. The figure 5.10 shows the attributes of this class. The tweet attribute keeps the tweet object. Boolean nvn attribute shows if the tweet matches the N-V-N pattern. If it matches, the nvn attribute is true, and otherwise it is false. The attributes wordOccMap, sigWordOccMap, nNounOccMap, verbOccMap are hashMaps and they keep all words, all words except stopwords, all nNoun and all verbs in the tweet with their occurrence frequencies, respectively. There is also an arrayList attribute called nvnList. If the tweet matches N-V-N pattern, the attributes of N-V-N sequence as first nNoun, verb and the second nNoun are stored in this arrayList. The attributes of the NvNTriplet object is shown in the figure 5.11.

In the pre-procces operation, it is determined if a tweet matches N-V-N pattern. For N-V-N matching, following operations are performed.

- The tweet is split into tokens.
- For each token:
  (i) Update TweetAnalysisData.wordOccMap.
  (ii) Remove punctuations. In this process all non-alpha characters are removed from the token.

```
TweetAnalysisData
(
Tweet tweet = null;
boolean nvn = false;
Map(String,Integer) wordOccMap = null;
Map(String,Integer) sigWordOccMap = null;
Map(String,Integer) nNounOccMap = null;
Map(String,Integer) verbOccMap = null;
ArrayList(NvNTriplet) nvnList = null;
)
```

Figure 5.10. The attributes of a TweetAnalysisData

```
NvNTriplet
(
String firstnNoun = null;
String verb = null;
String secondnNoun = null;
)
```

Figure 5.11. The attributes of a NvNTriplet Object

(iii) Remove the token if it is a stop word

(iv) If token is not a stop word update TweetAnalysisData.sigWordOccMap

(v) Stem the token

(vi) Get parts of speech of the token.

- If token is an nNoun update TweetAnalysisData.nWordOccMap.
- If token is a verb update TweetAnalysisData.verbOccMap.
- If tweet matches N-V-N pattern set TweetAnalysisData.nvn as true.

Figure 5.12 shows the code snippet of the pre-proccess operation. In the pre-processing operation:

- getPos() function returns part of speech of the given word. The word is first stemmed then WordNET is utilized while determining part-of-speech of the word. If WordNet cannot identify the word, it is assumed that the word is a candidate of nNoun and count as nNoun. Figure 5.13 shows the algorithm of getPos() function and figure 5.14 describes the posData object which is used by the getPos() function. The posData object keeps two booleans as nNoun and verb which are used while determining the N-V-N sequence.
- While searching N-V-N sequence in a tweet, the following methodology is used:
  (i) Traverse each token in the tweet.
  (ii) If an nNoun is found, check if another nNoun and a verb were found before. If found, the tweet has N-V-N sequence, if not, keep that a nNoun is found.
  (iii) If a verb is found, check if an nNoun was found before. If found, keep that a verb is found, if not, omit it.
  (iv) The pseudocode in figure 5.15 describes the operation of finding N-V-N sequence in detail.

In the post-process operation, the overall data is gathered from all tweets. In this operation the followings are obtained for a set of tweets.

- Total number news tweets.
- Total number of nNouns

```
TweetAnalysisData preProcessTweet(Tweet tweet)
(
TweetAnalysisData tweAnlysData = new TweetAnalysisData();
NvNTriplet nvnTriplet = new NvNTriplet();
tweAnalysisData.setTweet(emoTweet);
tokenizer st = new Tokenizer(tweet.getTweetText());
while st.hasMoreTokens() do
  String currentWord = st.nextToken();
  updateMap(tweAnlysData.getWordOccurenceHashMap());
  currentWord = clearPunctuation(currentWord);
  if currentWord!=null AND !isStopWord(currentWord) then
    updateMap(tweAnlysData.getSigWordOccurenceHashMap());
    if getPOS(currentWord).isnNoun() then
      updateMap(tweAnlysData.getnNounOccurenceHashMap());
      if NvNTriplet.getFirstnNoun()!=null AND NvNTriplet.getVerb()!=null
      then
        tweAnlysData.setNVN(true);
        updateNvNTriplet();
      else
        updateNvNTriplet();
      end if
    end if
    if getPOS(currentWord).isVerb() then
      updateMap(tweAnlysData.getVerbOccurenceHashMap());
      updateNvNTriplet();
    end if
  end if
end whileupdateNVNList(tweAnlysData);
)
```

Figure 5.12. The algorithm for tweet pre-proccesing

```
POSData getPos(String token)
(
POSData posData = new POSData();
if wordNet.getVerb(currentWord)!=null then
   posData.setVerb(true);
else
   if wordNet.getAdjective(currentWord)!=null then
      do nothing
   end if
else
   if wordNet.getAdverb(currentWord)!=null then
      do nothing
   end if
else
   if wordNet.getNoun(currentWord)!=null then
      posData.setnNoun(true);
   end if
else
   posData.setnNoun(true);
end if
return posData();
)
```

Figure 5.13. The algorithm for getPos() function

```
POSData
(
boolean nNoun = null;
boolean verb = null;
)
```

Figure 5.14. The attributes of a POSData object

```
String firstnNoun = NULL;
String verb = NULL;
String secondnNoun = NULL;
for all token in the tweet do
  if token is nNoun) then
    if firstnNoun is not NULL) then
      if verb is not NULL) then
        tweet contains N-V-N sequence;
        return true;
      end if
    else
      firstnNoun = token;
    end if
  end if
  if token is verb) then
    if firstnNoun is not NULL) then
      verb = token;
    end if
  end if
end for
```

Figure 5.15. The pseudocode for finding N-V-N sequence

- Total number of verbs

(ii) The **performAnalysis** method performs same operations what searchAndReturnResults does, except instead of searching Twitter with a given keyword, it makes analysis on the tweets already stored in the database. The performAnalysis method allows us to analyze more than 1500 tweets at a time.

5.1.2. Client Side Implementation

Client side of Twiggle was created utilizing Google Web Toolkit, it consists of the user interface of the Twiggle which is responsible for the following operations:

- Provides an interface to the end-user to insert a keyword (see section 4.3.1) and transmit this keyword to the server side using web service methods.
- Displays results to the users.



Figure 5.16. Twiggle Logo and Search Interface

The Twiggle logo and the main search interface is shown in figure 5.16. The user can simply insert a keyword to the text box and click on the search button. When clicking on the search button in the background searchAndReturnResults web service method is called.

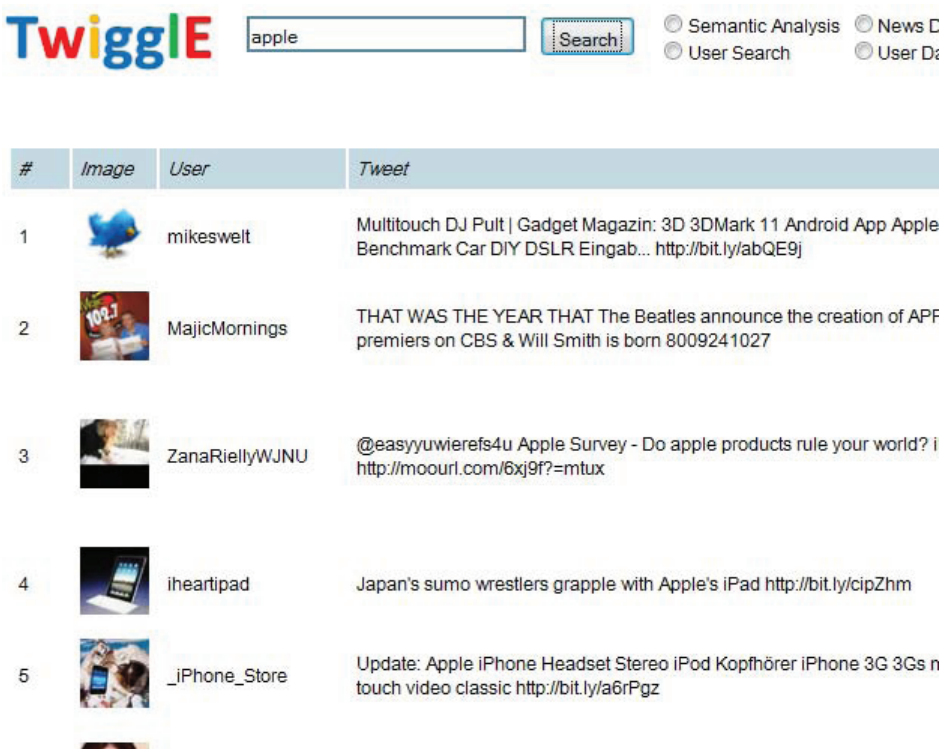


Figure 5.17. Twiggle Result Interface

Figure 5.17 shows how Twiggle displays the tweets which are set by the system as news tweets.

6. EXPERIMENTS AND RESULTS

In this chapter the case studies performed on the data collected from Twitter are described. In the studies, the news tweet is denoted as a tweet which match N-V-N pattern.

6.1. Case Study: User Analysis

In this section, the contributions of some selected Twitter users were analyzed. Table 6.1 shows the categories of the selected users based on their contributions. The users were chosen randomly and from WeFollow [33] which is an Internet site that categories most followed twitter users in terms of their interests. The users in the category news are the mass media which uses Twitter to spread news information. The users in the celebrity category are the famous people in real life. Since their popularity, they generally have millions of followers in Twitter. The users in the company category are the companies which uses Twitter to advertise their products or to give news about company itself. The users in the academic category utilize Twitter for academic purposes. There are also other Twitter users which are popular in [33] were put under misc category and randomly chosen users were put under category random. The individual records for every user are shown in appendix B.

Table 6.1. The categories of the user in terms of their interests

News	**Celebrity**	**Misc**	NiemanLab
cnnbrk	aplusk	Jason	leolaporte
bbcnews	Oprah	shoemoney	**Random**
TechCrunch	BarackObama	smashingmag	TeamMalachiae
TheEconomist	TheEllenShow	postsecret	mathrabbit1
nytimes	britneyspears	ijustine	greatspeaking
BreakingNews	THE_REAL_SHAQ	dsearls	stjohnk5
TIME	MariahCarey	hrheingold	orangeunicorns
mashable	50cent	davewiner	anna0974
CBSNews	snoopdogg	Mediabistro	iTauqeer
Poynter	PerezHilton	andersoncooper	LUKIKA
journalismnews	**Academic**	kevinrose	Alyssafelldown
Newsweek	zef	jayrosen_nyu	Hajji_love
TheOnion	BryanAlexander	GuyKawasaki	CallaLove
Company	EelcoVisser	Scobleizer	
google	johnbreslin	iamdiddy	
zappos	uskudarli	Veronica	

These twitter users were studied to expose how often they post news tweets. Figure 6.1 shows the list of the users ordered by the ratio of news tweets they posted.

Table 6.2: News tweet Ratio for selected users

#	**User Name**	**News Tweet Ratio**
1	BreakingNews	0,95
2	Poynter	0,93
3	bbcnews	0,93
4	NBCNews	0,90
5	BarackObama	0,83

Table 6.2 – continued

#	User Name	News Tweet Ratio
6	cnnbrk	0,81
7	postsecret	0,80
8	TheEconomist	0,80
9	journalismnews	0,79
10	iTauqeer	0,79
11	zappos	0,79
12	MariahCarey	0,78
13	CBSNews	0,77
14	jayrosen_nyu	0,77
15	NiemanLab	0,77
16	TechCrunch	0,73
17	BryanAlexander	0,71
18	Mediabistro	0,70
19	Oprah	0,68
20	britneyspears	0,68
21	TheOnion	0,67
22	google	0,67
23	stjohnk5	0,66
24	greatspeaking	0,66
25	TheEllenShow	0,65
26	Scobleizer	0,63
27	nytimes	0,63
28	mashable	0,63
29	leolaporte	0,62
30	EelcoVisser	0,61
31	Newsweek	0,60
32	iamdiddy	0,58
33	johnbreslin	0,56
34	mathrabbit1	0,55

Table 6.2 – continued

#	User Name	News Tweet Ratio
35	LUKIKA	0,54
36	hrheingold	0,54
37	dsearls	0,53
38	Hajji_love	0,52
39	snoopdogg	0,52
40	Jason	0,49
41	50cent	0,49
42	andersoncooper	0,46
43	zef	0,46
44	davewiner	0,45
45	ijustine	0,44
46	smashingmag	0,43
47	aplusk	0,41
48	anna0974	0,41
49	TIME	0,41
50	orangeunicorns	0,39
51	THE_REAL_SHAQ	0,38
52	TeamMalachiae	0,38
53	shoemoney	0,38
54	kevinrose	0,36
55	uskudarli	0,34
56	Veronica	0,29
57	PerezHilton	0,29
58	GuyKawasaki	0,27
59	CallaLove	0,26
60	Alyssafelldown	0,19

According to figure 6.1, it is observed that the 7 of top 10 most news poster are the users in the news category. These results show that, using N-V-N pattern to identify news tweet is encouraging. The followings indicate some interesting points of this analysis.

- 13 Twitter users were chosen in the news category and it is seen that 7 of them are among the top 10 most news poster in the list.
- BreakingNews is the user having the most news tweet ratio with 95 per cent. BreakingNews is a bot and spreads news from different media agencies. When the tweets of the BreakingNews are inspected, it is seen that almost every tweet consists of fully qualified sentence. Most of the time when the tweet is read, the overall news content can be understood easily. The following tweet is an example news tweet from BreakingNews where missiles, fire and embassy are noun-verb-noun and they form N-V-N sequence.

  Missiles fired at U.S. Embassy in Baghdad during VP Biden's visit http://bit.ly/a6Nicp

- Having 41 per cent news tweet ratio TIME takes the last place among the users in the news category. When the tweets that TIME sent were inspected, it is observed that, TIME is not a bot and does not only spread tweets, it also response other twitter users and most of the time its posts consist of unfinished sentences. The followings are the non-news tweet examples from the user TIME.

  @cksopher Ouch.
  @CallCarpenter Thanks!
  This week in web trends — http://su.pr/2UzfXG

- The user BarackObama represents the official Twitter account of the US president Barrack Obama. It is the 5th most news poster in the list. The tweets of the user BarackObama, consist of most of time fully qualified sentences with obeying rules of formal English. The followings are some example news tweets from user BarrackObama.

  History was made today when Sonia Sotomayor took the judicial oath and joined the Supreme Court. Congratulations, Justice.

  where Noun-Verb-Noun sequence formed with sotomayor,take,oath respectively.

  She's a trailblazing lawyer who has dedicated her life to public service. Today Elena Kagan began hearings to become a Supreme Ct. justice.

where Noun-Verb-Noun sequences are formed with kagan, begin, hearings and lawyer, dedicated, life respectively.

- The postsecret represents the Internet site Postsecret where people post in their secrets anonymously. The tweets of the user postsecret, generally gives secrets submitted to Postsecret Internet site and also news about the PostSecret Internet site itself. Example tweets are the followings:

  > An ASL Interpreter has been requested for the PostSecret Event in Baltimore this Sunday at noon. If you can help please tweet me.

  where Noun-Verb-Noun sequences are formed with interpreter, requested, postsecret respectively.

  > Today's Email: "On airplanes I wonder what role each passenger would play in survival if the plane crashed.

  where Noun-Verb-Noun sequences are formed with passenger, play, survival respectively.

- The user MariahCarey represents the official Twitter account of the famous pop singer Mariah Carey. She is in the 11th in the list with 78 per cent news tweet ratio. The tweets of user MariahCarey gives news about Mariah Carey. Example tweets from MariahCarey are the followings:

  > Tune in to 'House' on FOX tonight for the premiere of the new AT&T commercial feat. Mariah directed by Oscar nom. director Bennett Miller

  where Noun-Verb-Noun sequences are formed with mariah, direct, oscar respectively.

  > Mariah looked stunning on the red carpet at the Golden Globes in support of her fim Precious see pics @ http://tinyurl.com/yzz9mbv (MC.com)

  where Noun-Verb-Noun sequences are formed with mariah, stun, red respectively.

- The user bryanalexander who posts new tweets for academic purposes. He has a high news tweet rate as 71 per cent. Example tweets from bryanalexander are the followings:

  > Most med students would play computer games to learn, http://bit.ly /bhHMZF . Would the same go for pre-med undergrads?

  where Noun-Verb-Noun sequences are formed with students, play, computer respectively.

  > Educause looks at alternatives to course management systems, http://bit.ly /bUZkHa

  where Noun-Verb-Noun sequences are formed with educause, look, alternatives respectively.

- The user GuyKawasaki is one of the most famous twitter users. He has more than 200.000 followers. On the other hand, it is observed that, he has the 3rd least news tweet ratio with 27 per cent. The tweets of the GuyKawasaki may contain news but more likely to be consist of unfinished sentences which does not match N-V-N pattern.The followings are some selected tweets of GuyKawasaki.

  Ring box LCD screen http://u.nu/366te
  Funny paper bag masks http://idek.net/2yjW
  100 best beers http://tinyurl.com/27z4kng

- The user AlyssafelldownIn is in the bottom of the list. Her news tweet ratio is 19 per cent. The tweets of the AlyssafelldownIn mainly consist of chitchatting. The followings are some selected tweets of AlyssafelldownIn.

  Wow.
  cool
  :O
  I knowwww :)

6.2. Case Study: World Cup Tweets

FIFA World Cup is one of the most attractive global sport events, and the latest one took place in South Africa between June 11th and July 11th 2010. During this organization some games were followed to analyze news tweets about the followed game.

6.2.1. Argentina vs Germany

The quarter-final game between Argentina and Germany was played on July 3rd, 2010. In the game Germany beat Argentina with a score of 4-0.In this study, the keywords Germany, Argentina and Messi who is a famous Argentinian striker were given to the system between the dates July 1st, 2010 and July 8th, 2010 inclusive. During this period the tweets including these words were collected from Twitter. The overall data collected for the keywords Messi, Argentina and Germany is shown in the tables C.5 , C.7 and C.3 respectively, in Appendix C.

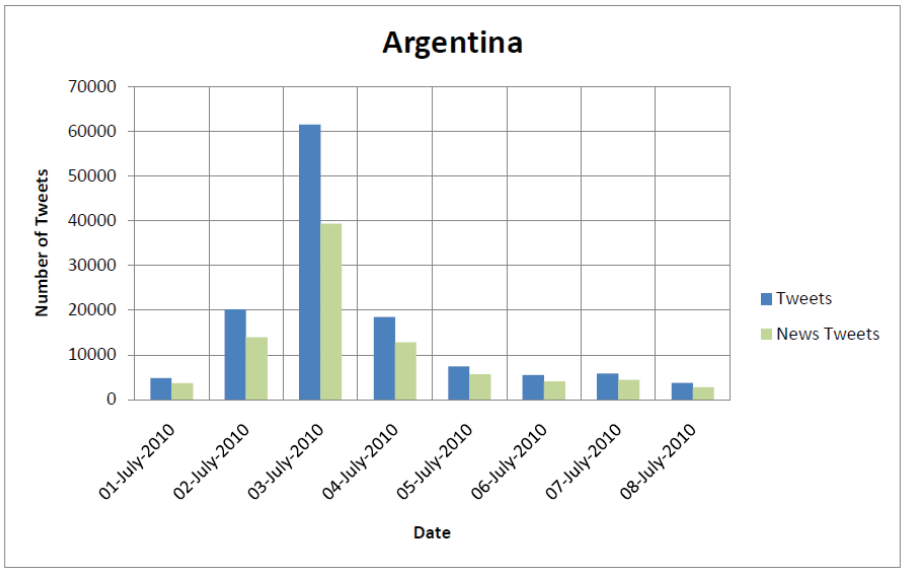


Figure 6.1. The distributions of all tweets and news tweets by days for keyword Argentina

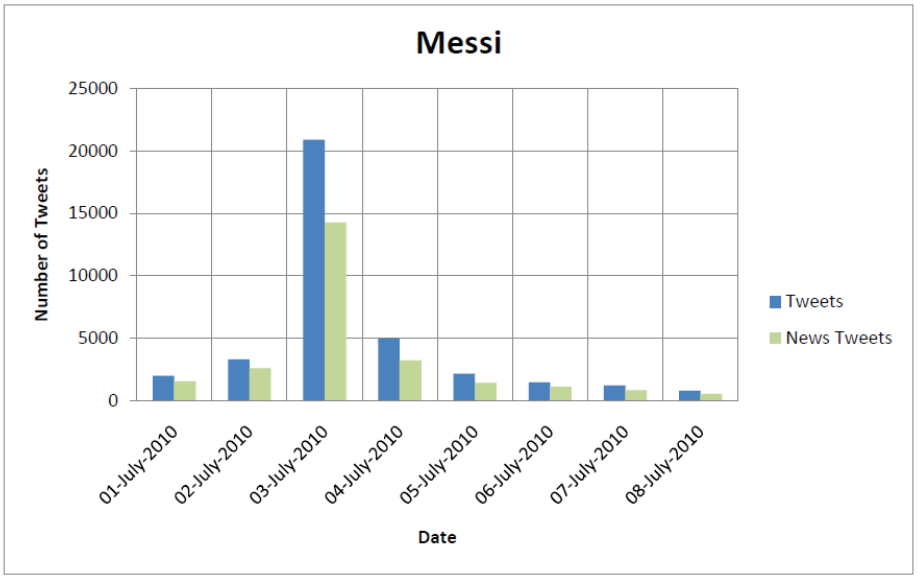


Figure 6.2. The distributions of all tweets and news tweets by days for keyword Messi

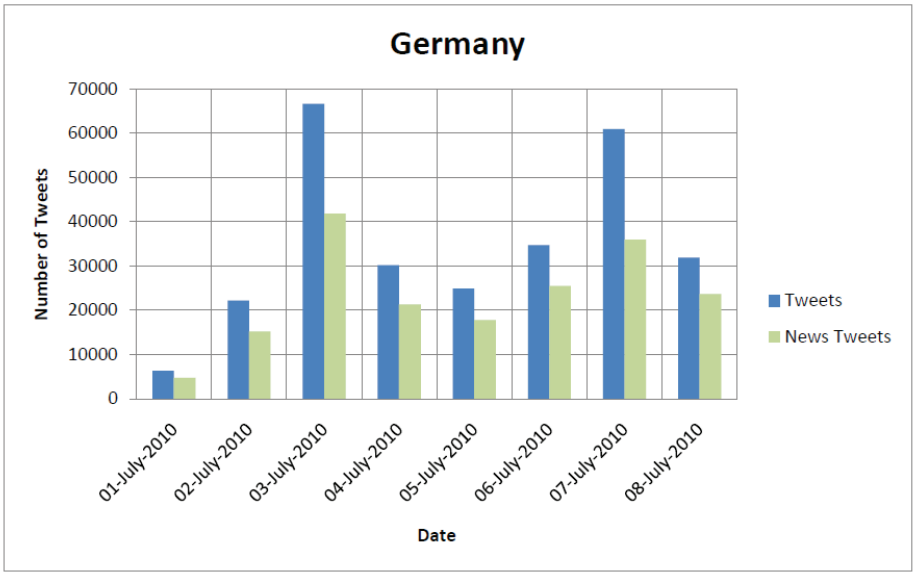


Figure 6.3. The distributions of all tweets and news tweets by days for keyword Germany

In the first analysis, news tweets over all the tweets are inspected daily for the keywords Messi, Argentina and Germany. Figures, 6.1, 6.2 and 6.3 shows the the results of the analysis for the keywords Argentina, Messi and Germany, respectively.

Two observations are gained from these results.

(i) The most tweets and the news tweets are posted on the day of the match, on July 3rd, 2010, when the Argentina vs Germany quarter final game was played.

(ii) As table 6.3 shows, the news tweets per tweets ratio is lowest on the match day for keywords Germany and Argentina. The reason is that, during matches people post tweets with emotionally filled and un-organized sentences with repetition of letters, capitalizations and exclamation marks.

AEEEEEEEEEEE ARGENTINA!
Argentina stillllllllllll
Chuuuuuuuuuupa Argentina!!!!!!!!! #worldcup
GO GERMANY!!!!!!!!
YESH GERMANY YESSSHHHHH

Table 6.3. The news tweets per tweets ratio for keywords Argentina, Messi and Germany

Date	**Argentina**	**Messi**	**Germany**
01-July-2010	0,77	0,78	0,75
02-July-2010	0,69	0,79	0,68
03-July-2010	0,64	0,68	0,63
04-July-2010	0,69	0,65	0,71
05-July-2010	0,76	0,66	0,72
06-July-2010	0,74	0,76	0,73
07-July-2010	0,76	0,69	0,59
08-July-2010	0,73	0,70	0,74

The figures 6.4, 6.5, 6.6, 6.7, 6.8, 6.9 show the occurrence frequencies of all nNouns found in the all tweets and news tweets. Since the keywords argentina, germany and messi are already exist in the retrieved tweets they are omitted from list. Based on these results it can be observed that, the Twitter users used some words more frequently than the others. These words most of the time, give clues about the news topic of the day of the tweets were submitted. For example, some most frequently occurred nNouns in 6.7 as Germany, Messi, football and semifinals indicates an existence of an football match.

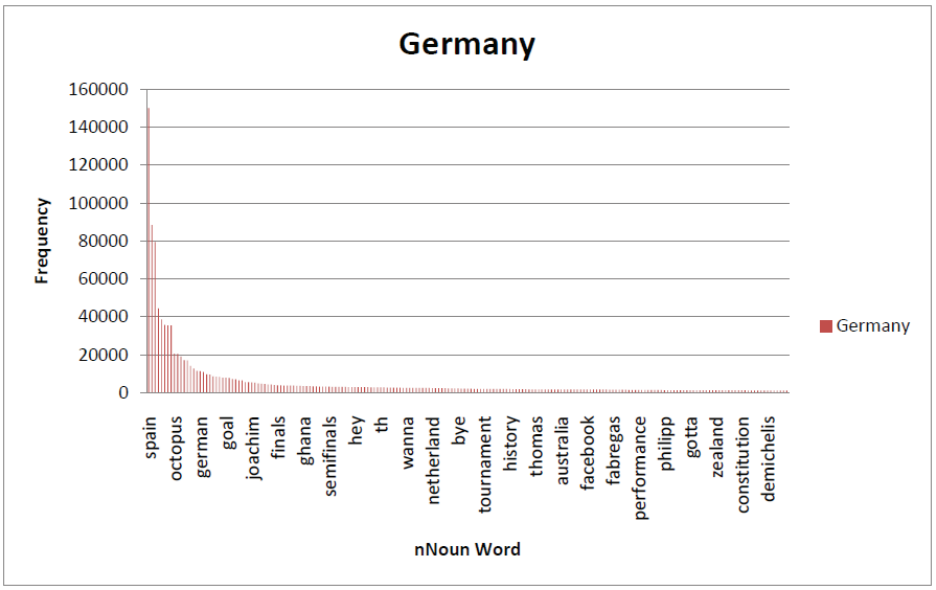


Figure 6.4. The occurrence frequency of all nNouns in all tweets for keyword Germany

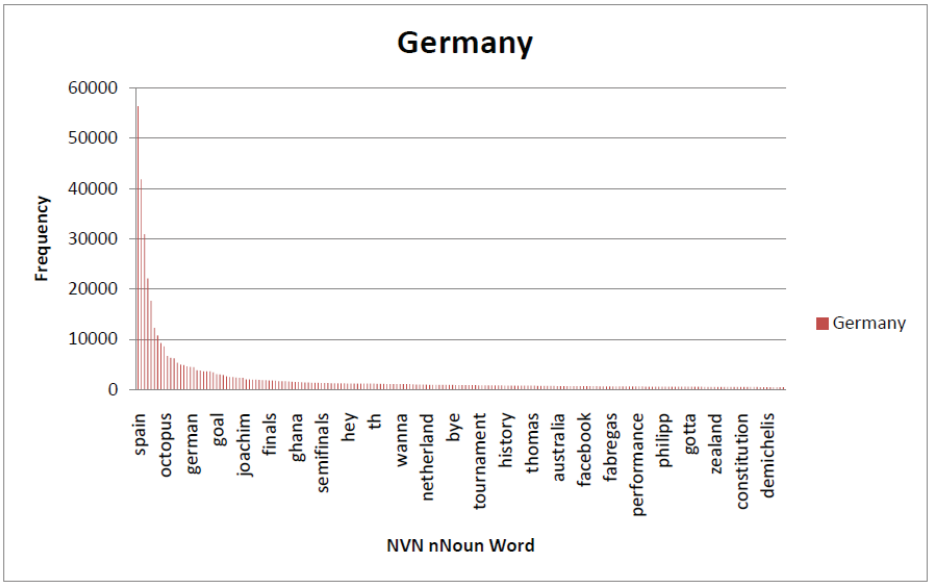


Figure 6.5. The occurrence frequency of all nNouns in news tweets for keyword Germany

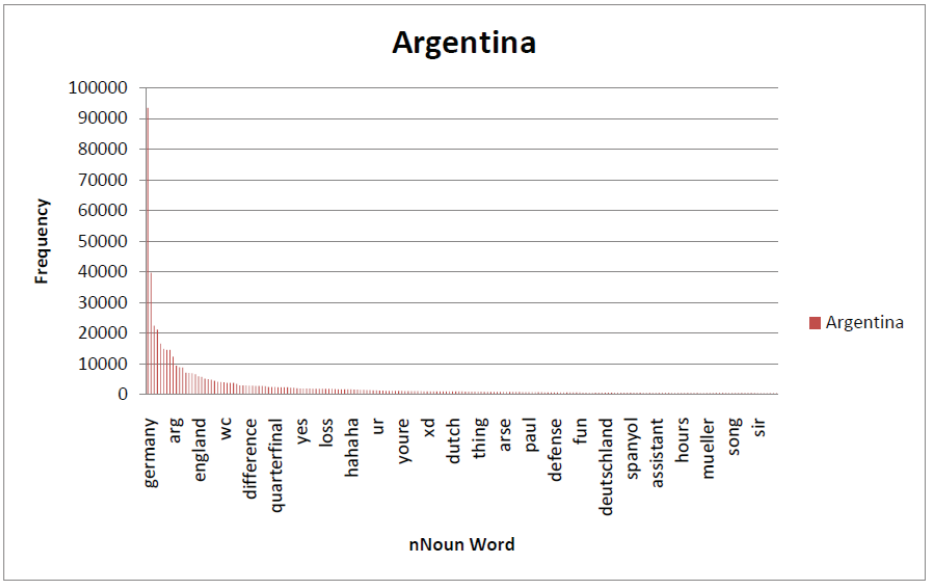


Figure 6.6. The occurrence frequency of all nNouns in all tweets for keyword Argentina

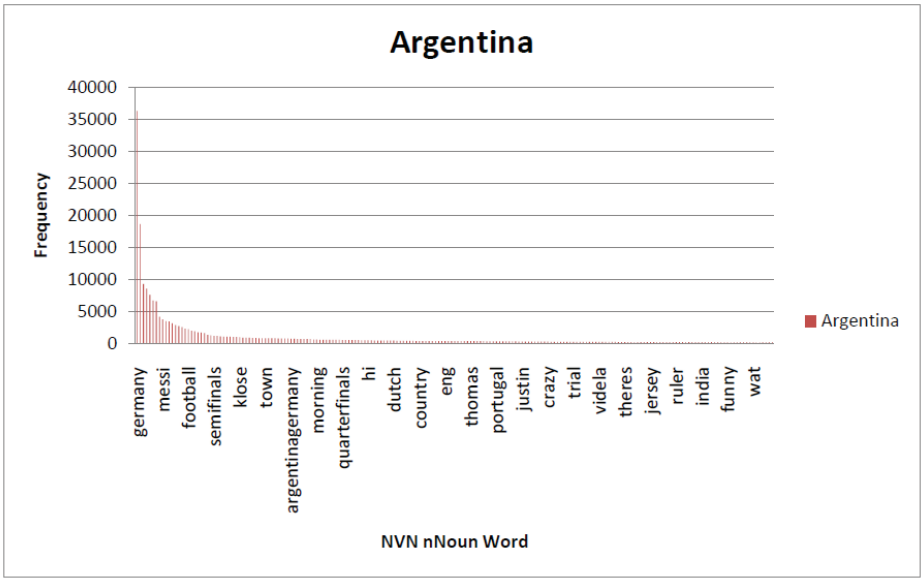


Figure 6.7. The occurrence frequency of all nNouns in news tweets for keyword Argentina

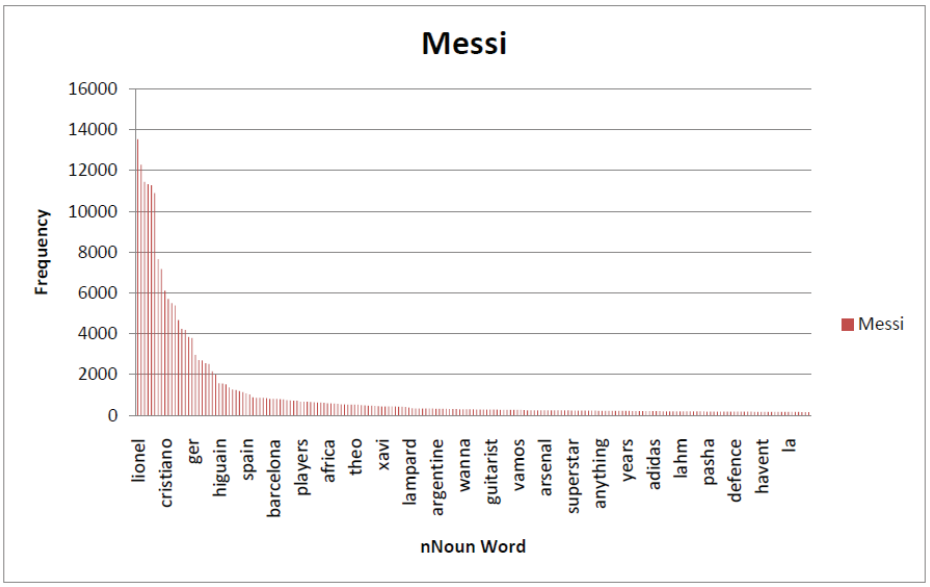


Figure 6.8. The occurrence frequency of all nNouns in all tweets for keyword Messi

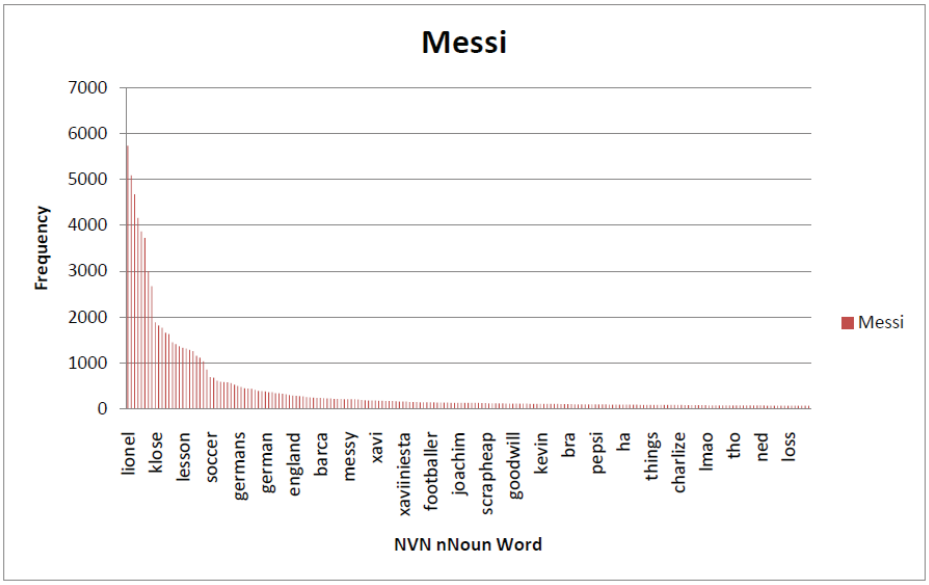


Figure 6.9. The occurrence frequency of all nNouns in news tweets for keyword Messi

6.2.2. Brazil

Brazil played Netherlands in the quarterfinals on 2nd of July and lost the game with a score of 2-1. Same as the first case study, the keyword Brazil, was given to the system between the dates July 1st, 2010 and July 8th, 2010 inclusive and the tweets including Brazil was collected from Twitter. Figure 6.10 shows that daily graph of change in the number of the news tweets and all tweets for keyword brazil. Table 6.4 indicates that in the match day the news tweet ratio is one of the lowest amongst all. The overall data collected for the keywords Brazil, is shown in the tables C.10 in Appendix C.

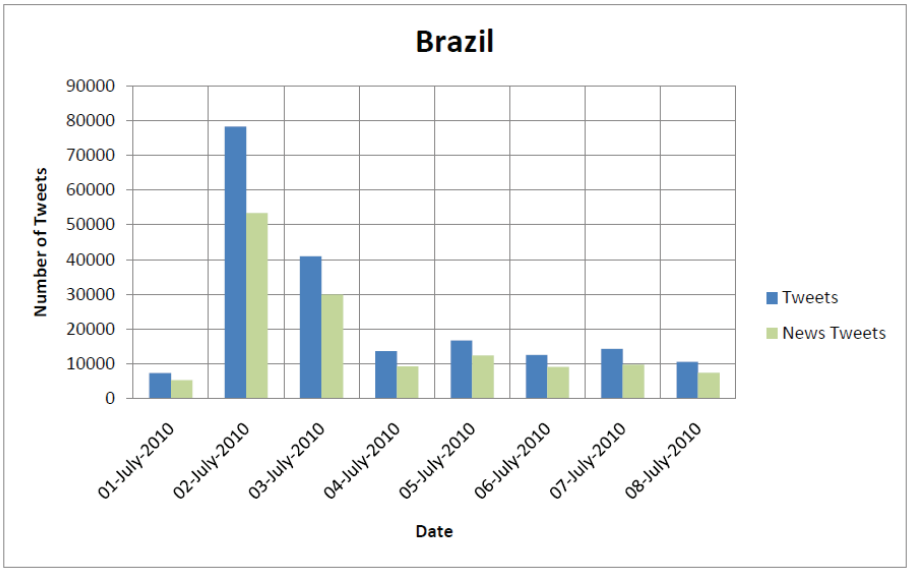


Figure 6.10. The distributions of all tweets and news tweets by days for keyword Brazil

Table 6.4. The news tweets per tweets ratio for keyword Brazil

Date	**Brazil**
01-July-2010	0,73
02-July-2010	0,68
03-July-2010	0,73
04-July-2010	0,68
05-July-2010	0,74
06-July-2010	0,72
07-July-2010	0,68
08-July-2010	0,70

6.2.3. Netherlands vs Spain

The World Cup final was played between Netherlands and Spain on 11th of July. Spain beat Netherlands with a score of 1-0 and won the World Cup. Spain and Netherlands keywords were given to the system between the dates July 11th, 2010 and July 12th, 2010 inclusive and the tweets including Spain and Netherlands were collected from Twitter. Figures 6.11 and 6.12 show that daily graph of change in the number of the news tweets and all tweets for keywords Spain and Netherlands. It is seen that, since Spain won the World Cup championship, the number of news tweets were posted on 12th of July was more than the day they won the championship. On the other hand, Netherlands has lower news post, on the day after they lost championship. The overall data collected for the keywords Netherlands and Spain is shown in the tables C.9 and C.2 respectively, in Appendix C.

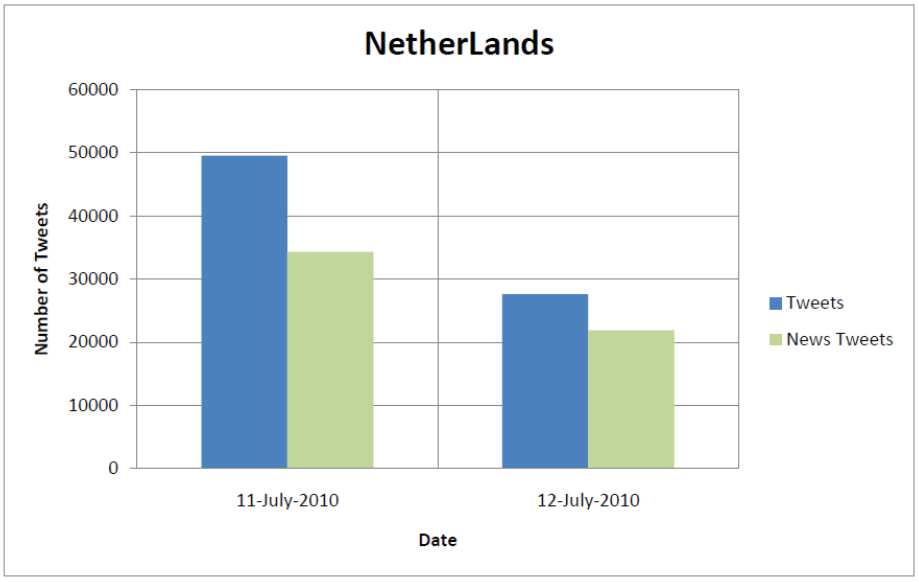


Figure 6.11. The distributions of all tweets and news tweets by days for keyword Netherlands

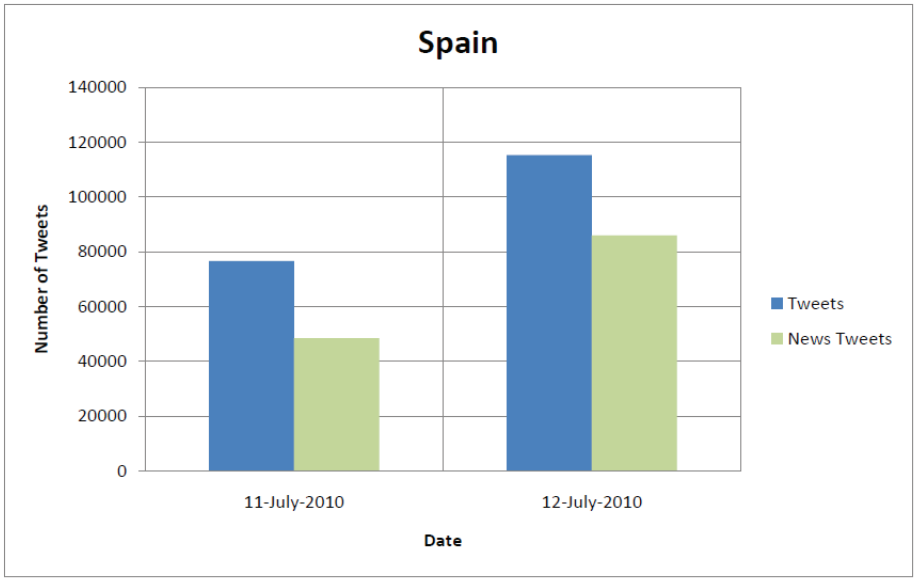


Figure 6.12. The distributions of all tweets and news tweets by days for keyword Spain

6.3. Case Study: Miscellaneous News

In this section miscellaneous news events were analyzed.

6.3.1. The Attack of Israel to Flotilla

Israel attacked Flotilla on 31st of May, 2010.In this study, the keywords flotilla, Gaza and Israel were given to the system between the dates May 31st, 2010 and June 5th, 2010 inclusive. During this period the tweets including these words were collected from Twitter. The overall data collected for the keywords Flotilla, Gaza and Israel is shown in the tables C.8 , C.11 and C.6 respectively, in Appendix C.

Firstly, news tweets over all the tweets are inspected daily for the keywords Flotilla, Gaza and Israel. Figures, 6.13, 6.14 and 6.15 shows the results of the analysis for the keywords Flotilla, Gaza and Israel, respectively.

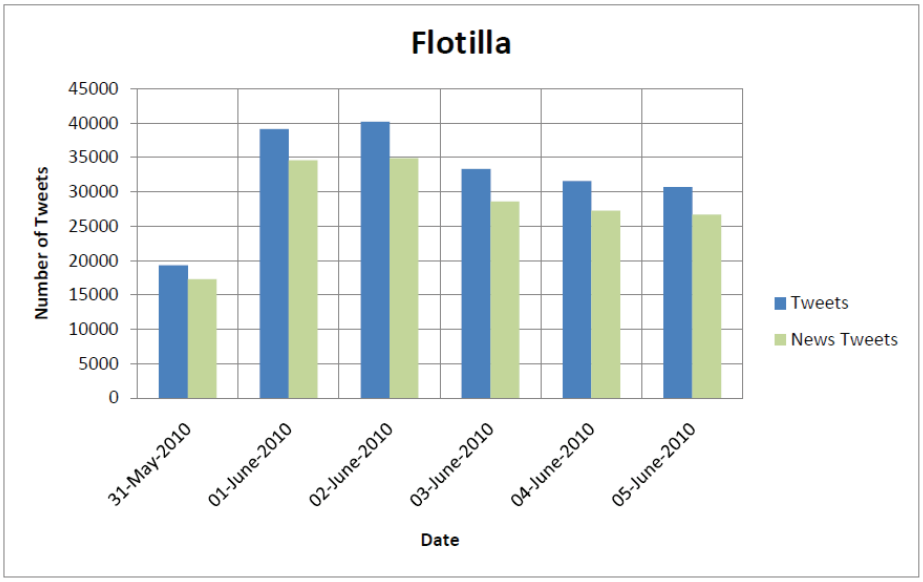


Figure 6.13. The distributions of all tweets and news tweets by days for keyword Flotilla

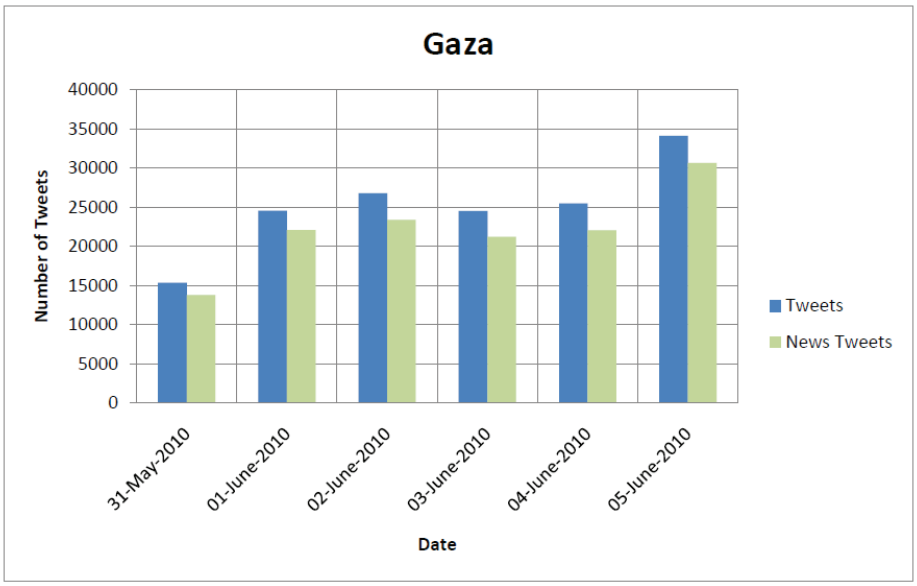


Figure 6.14. The distributions of all tweets and news tweets by days for keyword Gaza

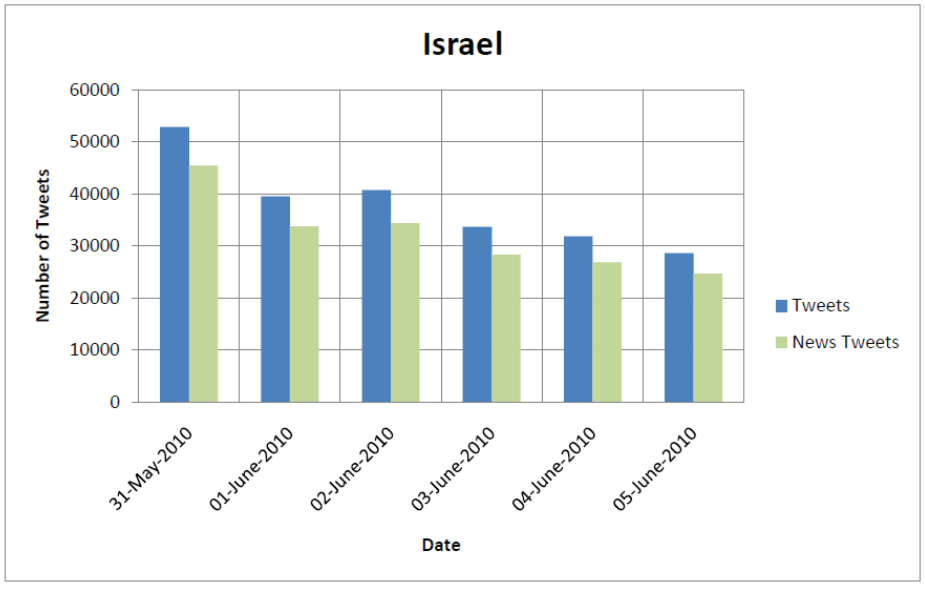


Figure 6.15. The distributions of all tweets and news tweets by days for keyword Israel

The followings are the observations gained from these results.

(i) Unlike sport events, the news tweets ratios in this event are closer to each other for each day. Additionally these ratios are higher than the news tweets ratios in sports event

(ii) As table 6.5 shows, the biggest gaps between news tweet ratios are 3 for Gaza and Israel and 4 for Flotilla.

Table 6.5. The news tweets per tweets ratio for keywords Flotilla, Gaza and Israel

Date	**Flotilla**	**Gaza**	**Israel**
01-July-2010	0,90	0,90	0,86
02-July-2010	0,88	0,90	0,85
03-July-2010	0,87	0,87	0,84
04-July-2010	0,86	0,87	0,84
05-July-2010	0,86	0,87	0,84
06-July-2010	0,87	0,90	0,86

6.3.2. Moscow Metro Bombing

The bombing event in Moscow metro happened on 29th of March, 2010.In this study, Moscow Metro keyword was given to the system between the dates March 29th, 2010 and March 31th, 2010 inclusive. During this period the tweets including Moscow Metro were collected from Twitter. The overall data collected for the keyword Moscow metro is shown in the table C.1 in Appendix C.

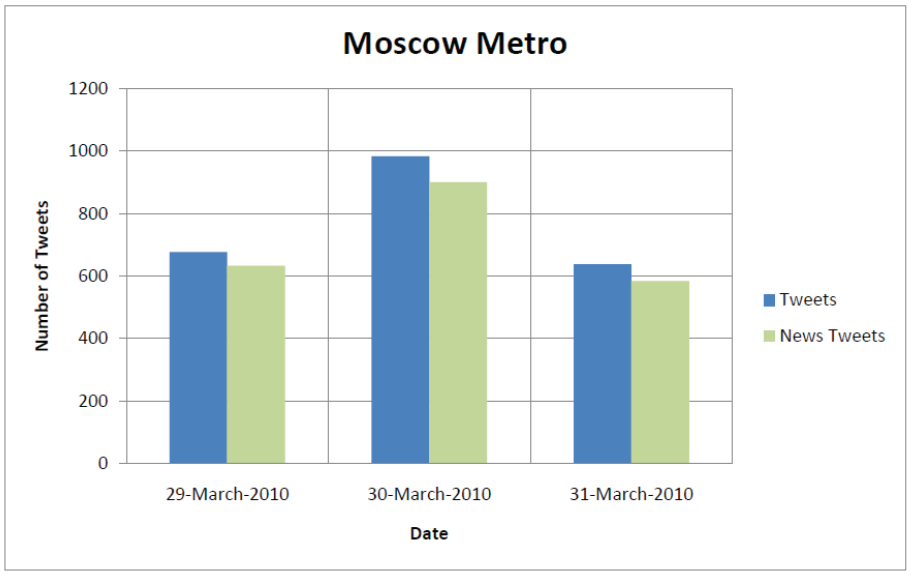


Figure 6.16. The distributions of all tweets and news tweets by days for keyword MoscowMetro

The results of the figure 6.16 supports the findings in the first section.

6.3.3. Oil spill disaster in Gulf of Mexico

The Oil spill disaster in Gulf of Mexico occurred on 20th of April, 2010.In this study, Oil spill keyword was given to the system between the dates May 23rd, 2010 and June 07th, 2010 inclusive. During this period the tweets including Oil spill were collected from Twitter. The overall data collected for the keyword Oil spill is shown in the table C.4 in Appendix C.

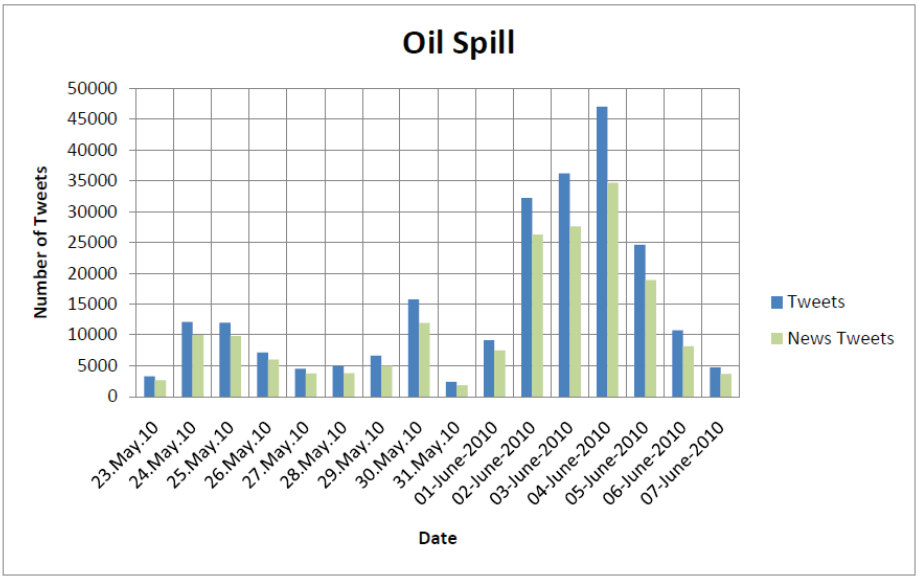


Figure 6.17. The distributions of all tweets and news tweets by days for keyword OilSpill

The figure 6.17 shows daily news tweet ratios during the days between May 23rd, 2010 and June 07th, 2010 for keyword Oil Spill. Oil spill event is a long lasting event and the fluctuations in the results indicate that, in some days the oil spill event is tweeted more than others.

6.3.4. Other Events

The tweets for the following keywords were also collected from twitter to make news observations but they are not analyzed. The overall data for them can be found in [34].

- april 1: gathered to get response from people about april1 day
- arsenal: gathered for the matches between Arsenal vs Barcelona on March 31th and April 6th
- barcelona gathered for the matches between Arsenal vs Barcelona on March 31th and April 6th

- barca: gathered for the matches between Arsenal vs Barcelona on March 31th and April 6th
- BBQKALBI: To retrieve local interest of the people about street food cart.(BBQKALBI is a Street food cart in San Francisco)
- Dennis Hopper: Dennis Hopper is an American actor, filmmaker and artist. Died in May 29th.
- Eartquake: Earthquake in California on April 4th, and earthquake in china on April 14th.
- Gary Coleman: Gary Coleman was an American actor.Died in May 28th.
- ist2010: Interferenze Seeds conference in Tokyo 2010 .
- istanbul: gathered to get general news from Istanbul.
- eurovision: eurovision on May 30th.
- iste2010: ISTE 2010 conference in Denver.
- Jose Lima: Jose Lima was an Baseball player.Died in May 23th.
- eurovision: eurovision on May 30th.
- Last Airbender: Last Airbender is a film.
- michael jackson: First anniversary day of the death of Michael Jackson.
- MoGoBBQ: To retrieve local interest of the people about street food cart.(MoGoBBQ is a Street food cart in San Francisco).
- New York: gathered to get general news from New York.
- Semantic: gathered to get general news from Semantic.
- sonisphere: Concert in İstanbul on Jun 25th-26th-27th.
- toronto conference: gathered to get general news from Conference in Toronto .
- volcano: gathered to get general news from Volcano eruption in Iceland on April 14th.
- vuvuzela: Popular music entstrument in the world cup.
- W3C: World Wide Web Consortium conference.
- world cup: General news tweets about World Cup.
- wwdc: Apple Worldwide Developers Conference on June 7th.

7. CONCLUSIONS

In this thesis, news analysis on microblog systems was performed. An approach was proposed to extract news contributions from microblogs and this approach was used in the experiments.

Twitter [4] was selected as the microblog system and as section 4.2 describes a pattern called N-V-N was utilized to extract news tweets from Twitter.

The experiments in chapter 6 were carried out in the following two categories:

(i) User tweets Analysis : In this experiment the tweets of the 60 users selected from Twitter were analyzed. The tweets of every user were analyzed individually to check if the tweets match N-V-N pattern. The users were ranked according to their news tweet ratios and a list was formed as shown in figure 6.1. The results were encouraging to use N-V-N pattern for extraction of news tweets in microblogs, since the users which describe themselves as news contributors take places in the top of the list.

(ii) Event Analysis : In this experiment, some global events were followed and numerous tweets were analyzed. According to the results of the experiments it is observed that:

- In the sport events like football matches people tend to post more tweets in the match day than the other days. On the other hand, in match day, news tweet ratios are lowest because most of the time emotionally filled but non-news tweets are posted.
- When number of nouns or named entities were analyzed in news tweets, it is seen that some word are much more used than the other. This tells us that, in news tweets people focus on some nouns than the others.

In the future work and discussions section 7.1 the possible options are discussed to improve results.

7.1. Future Work and Discussions

The following can be thought as future works:

- Improvement on news pattern : The suggested news pattern in this thesis was N-V-N pattern which is heavily rely on wordNet while deciding nouns and verbs. Since a word can be noun in one post but verb in another, deciding accurate part of speech of the word in a post was a problem. In this thesis work, if a word has a verb and noun meaning, its pos was assumed as verb ignoring its noun meaning. As an improvement, we can check both noun and verb meaning while searching N-V-N sequence in a post, so that we can increase the possibility of existence of an N-V-N sequence in a post. On the other hand, semantic analysis or NLP can also be utilized to determine the part of speech of the word in a post. By this way the accuracy of the N-V-N pattern matching will be improved and better results will be gained.
- Another news pattern : In this work only one news pattern was examined, other patterns can be searched and applied on microblogs. The results can be compared with N-V-N pattern to perceive advantages and disadvantages of news patterns over each other.
- Different news events : In this thesis, the news analysis performed in the case studies, cover global events as football matches in the world cup and events which are mainly discussed in mass media. Other news areas such as news in local events which do not spread through mass media can also be studied. This type of work could be useful to find out if the news about such events spread via a news pattern.
- Emotion Analysis : Emotion analysis can be performed on the news tweets. Some syntactic properties like repetition and capitalization can be utilized to comprehend in what emotions the news tweets are posted. As a result, overall emotional responses of the microblog users to a news event can be achieved.

APPENDIX A: STOP WORDS LIST

Stop words are the words frequently used in a natural language which do not have distinguishing property. In this study the stop words list is taken from [35].

The stop word list is the following: a's, able, about, above, according, accordingly, across, actually, after, afterwards, again, against, ain't, all, allow, allows, almost, alone, along, already, also, although, always, am, among, amongst, an, and, another, any, anybody, anyhow, anyone, anything, anyway, anyways, anywhere, apart, appear, appreciate, appropriate, are, aren't, around, as, aside, ask, asking, associated, at, available, away, awfully, be, became, because, become, becomes, becoming, been, before, beforehand, behind, being, believe, below, beside, besides, best, better, between, beyond, both, brief, but, by, c'mon, c's, came, can, can't, cannot, cant, cause, causes, certain, certainly, changes, clearly, co, com, come, comes, concerning, consequently, consider, considering, contain, containing, contains, corresponding, could, couldn't, course, currently, definitely, described, despite, did, didn't, different, do, does, doesn't, doing, don't, done, down, downwards, during, each, edu, eg, eight, either, else, elsewhere, enough, entirely, especially, et, etc, even, ever, every, everybody, everyone, everything, everywhere, ex, exactly, example, except, far, few, fifth, first, five, followed, following, follows, for, former, formerly, forth, four, from, further, furthermore, get, gets, getting, given, gives, go, goes, going, gone, got, gotten, greetings, had, hadn't, happens, hardly, has, hasn't, have, haven't, having, he, he's, hello, help, hence, her, here, here's, hereafter, hereby, herein, hereupon, hers, herself, hi, him, himself, his, hither, hopefully, how, howbeit, however, i'd, i'll, i'm, i've, ie, if, ignored, immediate, in, inasmuch, inc, indeed, indicate, indicated, indicates, inner, insofar, instead, into, inward, is, isn't, it, it'd, it'll, it's, its, itself, just, keep, keeps, kept, know, knows, known, last, lately, later, latter, latterly, least, less, lest, let, let's, like, liked, likely, little, look, looking, looks, ltd, mainly, many, may, maybe, me, mean, meanwhile, merely, might, more, moreover, most, mostly, much, must, my, myself, name, namely, nd, near, nearly, necessary, need, needs, neither, never, nevertheless, new, next, nine, no, nobody, non, none, noone, nor, normally, not, nothing, novel, now, nowhere, obviously, of, off, often,

oh, ok, okay, old, on, once, one, ones, only, onto, or, other, others, otherwise, ought, our, ours, ourselves, out, outside, over, overall, own, particular, particularly per, perhaps, placed, please, plus, possible, presumably, probably, provides, que, quite, qv, rather, rd, re, really, reasonably, regarding, regardless, regards, relatively, respectively, right, said, same, saw, say, saying, says, second, secondly, see, seeing, seem, seemed, seeming, seems, seen, self, selves, sensible, sent, serious, seriously, seven, several, shall, she, should, shouldn't, since, six, so, some, somebody, somehow, someone, something, sometime, sometimes, somewhat, somewhere, soon, sorry, specified, specify, specifying, still, sub, such, sup, sure, t's, take, taken, tell, tends, th, than, thank, thanks, thanx, that, that's, thats, the, their, theirs, them, themselves, then, thence, there, there's, thereafter, thereby, therefore, therein, theres, thereupon, these, they, they'd, they'll, they're, they've, think, third, this, thorough, thoroughly, those, though, three, through, throughout, thru, thus, to, together, too, took, toward, towards, tried, tries, truly, try, trying, twice, two, un, under, unfortunately, unless, unlikely, until, unto, up, upon, us, use, used, useful, uses, using, usually, value, various, very, via, viz, vs, want, wants, was, wasn't, way, we, we'd, we'll, we're, we've, welcome, well, went, were, weren't, what, what's, whatever, when, whence, whenever, where, where's, whereafter, whereas, whereby, wherein, whereupon, wherever, whether, which, while, whither, who, who's, whoever, whole, whom, whose, why, will, willing, wish, with, within, without, won't, wonder, would, would, wouldn't, yes, yet, you, you'd, you'll, you're, you've, your, yours, yourself, yourselves, zero.

APPENDIX B: TWITTER USER DATA

Table B.1. The summary of the collected data for user MariahCarey

UserName	MariahCarey		
Start Date	2009-01-28		
End Date	2010-08-17		
	All	**NvN**	**NvN/All**
# of tweets	794	622	0,78
# of dis. tweets	3.243	2.991	0,92
# of words	13.821	12.048	0,87
# of dis. words	3.243	2.991	0,92
# of sign. words	7.603	6.728	0,88
# of dis. sign. words	2.879	2.647	0,92
# of nNoun	3.753	3.348	0,89
# of dist. nNoun	1.862	1.576	0,85
# of verbs	2.963	2.632	0,89
# of dist. verbs	948	891	0,94
# of RTs	1	1	1

Table B.2. The summary of the collected data for user NBCNews

UserName	NBCNews		
Start Date	2008-09-03		
End Date	2010-08-18		
	All	**NvN**	**NvN/All**
# of tweets	3.026	2.725	0,9
# of dis. tweets	12.351	11.778	0,95
# of words	48.884	45.655	0,93
# of dis. words	12.351	11.778	0,95
# of sign. words	32.019	30.064	0,94
# of dis. sign. words	9.185	8.883	0,97
# of nNoun	16.036	15.343	0,96
# of dist. nNoun	5.623	4.965	0,88
# of verbs	13.649	12.612	0,92
# of dist. verbs	3.513	3.384	0,96
# of RTs	108	79	0,73

Table B.3. The summary of the collected data for user hrheingold

UserName	hrheingold		
Start Date	2009-12-10		
End Date	2010-08-19		
	All	**NvN**	**NvN/All**
# of tweets	2.910	1.561	0,54
# of dis. tweets	7.184	5.821	0,81
# of words	38.334	25.438	0,66
# of dis. words	7.184	5.821	0,81
# of sign. words	20.551	14.252	0,69
# of dis. sign. words	6.346	5.253	0,83
# of nNoun	8.444	6.396	0,76
# of dist. nNoun	3.649	2.742	0,75
# of verbs	9.504	6.228	0,66
# of dist. verbs	2.457	2.044	0,83
# of RTs	34	22	0,65

Table B.4. The summary of the collected data for user Mediabistro

UserName	Mediabistro		
Start Date	2009-06-17		
End Date	2010-08-18		
	All	**NvN**	**NvN/All**
# of tweets	3.182	2.223	0,7
# of dis. tweets	8.802	7.251	0,82
# of words	45.315	33.326	0,74
# of dis. words	8.802	7.251	0,82
# of sign. words	25.494	19.594	0,77
# of dis. sign. words	6.193	5.363	0,87
# of nNoun	11.098	9.180	0,83
# of dist. nNoun	3.720	2.993	0,8
# of verbs	11.622	8.459	0,73
# of dist. verbs	2.323	1.987	0,86
# of RTs	157	115	0,73

Table B.5. The summary of the collected data for user mashable

UserName	mashable		
Start Date	2010-04-18		
End Date	2010-08-19		
	All	**NvN**	**NvN/All**
# of tweets	3.197	2.003	0,63
# of dis. tweets	7.454	5.567	0,75
# of words	29.600	19.424	0,66
# of dis. words	7.454	5.567	0,75
# of sign. words	18.611	12.748	0,68
# of dis. sign. words	4.341	3.544	0,82
# of nNoun	8.932	6.598	0,74
# of dist. nNoun	2.470	1.826	0,74
# of verbs	7.867	5.158	0,66
# of dist. verbs	1.852	1.514	0,82
# of RTs	0	0	0

Table B.6. The summary of the collected data for user shoemoney

UserName	shoemoney		
Start Date	2009-04-17		
End Date	2010-08-19		
	All	**NvN**	**NvN/All**
# of tweets	3.161	1.192	0,38
# of dis. tweets	5.747	3.997	0,7
# of words	33.528	17.617	0,53
# of dis. words	5.747	3.997	0,7
# of sign. words	18.102	9.898	0,55
# of dis. sign. words	4.707	3.430	0,73
# of nNoun	6.714	4.192	0,62
# of dist. nNoun	2.761	1.863	0,67
# of verbs	8.665	4.484	0,52
# of dist. verbs	1.887	1.343	0,71
# of RTs	81	53	0,65

Table B.7. The summary of the collected data for user Jason

UserName	Jason		
Start Date	2010-04-13		
End Date	2010-08-19		
	All	**NvN**	**NvN/All**
# of tweets	2.704	1.329	0,49
# of dis. tweets	6.402	5.003	0,78
# of words	31.349	20.717	0,66
# of dis. words	6.402	5.003	0,78
# of sign. words	17.266	11.869	0,69
# of dis. sign. words	5.148	4.194	0,81
# of nNoun	8.146	5.741	0,7
# of dist. nNoun	3.126	2.305	0,74
# of verbs	7.507	5.144	0,69
# of dist. verbs	1.944	1.622	0,83
# of RTs	15	9	0,6

Table B.8. The summary of the collected data for user Oprah

UserName	Oprah		
Start Date	2009-04-17		
End Date	2010-05-20		
	All	**NvN**	**NvN/All**
# of tweets	98	67	0,68
# of dis. tweets	705	582	0,83
# of words	1.497	1.109	0,74
# of dis. words	705	582	0,83
# of sign. words	810	626	0,77
# of dis. sign. words	552	454	0,82
# of nNoun	355	297	0,84
# of dist. nNoun	301	229	0,76
# of verbs	382	279	0,73
# of dist. verbs	251	201	0,8
# of RTs	5	2	0,4

Table B.9. The summary of the collected data for user TheEllenShow

UserName	TheEllenShow		
Start Date	2009-03-10		
End Date	2010-08-19		
	All	**NvN**	**NvN/All**
# of tweets	2.536	1.655	0,65
# of dis. tweets	6.363	5.173	0,81
# of words	42.882	30.393	0,71
# of dis. words	6.363	5.173	0,81
# of sign. words	20.782	15.374	0,74
# of dis. sign. words	4.755	4.090	0,86
# of nNoun	8.430	6.735	0,8
# of dist. nNoun	2.817	2.241	0,8
# of verbs	9.720	6.897	0,71
# of dist. verbs	1.862	1.594	0,86
# of RTs	7	7	1

Table B.10. The summary of the collected data for user GuyKawasaki

UserName	GuyKawasaki		
Start Date	2010-07-30		
End Date	2010-08-19		
	All	**NvN**	**NvN/All**
# of tweets	3.197	852	0,27
# of dis. tweets	5.053	2.360	0,47
# of words	26.130	9.137	0,35
# of dis. words	5.053	2.360	0,47
# of sign. words	13.829	5.315	0,38
# of dis. sign. words	2.862	1.590	0,56
# of nNoun	5.417	2.625	0,48
# of dist. nNoun	1.478	782	0,53
# of verbs	6.871	2.257	0,33
# of dist. verbs	1.297	683	0,53
# of RTs	11	10	0,91

Table B.11. The summary of the collected data for user journalismnews

UserName	journalismnews		
Start Date	2009-11-27		
End Date	2010-08-19		
	All	**NvN**	**NvN/All**
# of tweets	3.185	2.530	0,79
# of dis. tweets	8.430	7.409	0,88
# of words	36.881	30.624	0,83
# of dis. words	8.430	7.409	0,88
# of sign. words	23.871	20.298	0,85
# of dis. sign. words	5.530	4.961	0,9
# of nNoun	13.293	11.143	0,84
# of dist. nNoun	3.393	2.730	0,8
# of verbs	8.469	7.450	0,88
# of dist. verbs	2.044	1.908	0,93
# of RTs	159	105	0,66

Table B.12. The summary of the collected data for user bbcnews

UserName	bbcnews		
Start Date	2010-07-20		
End Date	2010-08-19		
	All	**NvN**	**NvN/All**
# of tweets	3.196	2.967	0,93
# of dis. tweets	12.404	11.900	0,96
# of words	57.490	53.828	0,94
# of dis. words	12.404	11.900	0,96
# of sign. words	32.576	30.686	0,94
# of dis. sign. words	8.983	8.711	0,97
# of nNoun	15.248	14.639	0,96
# of dist. nNoun	5.661	5.009	0,88
# of verbs	14.657	13.623	0,93
# of dist. verbs	3.242	3.141	0,97
# of RTs	0	0	0

Table B.13. The summary of the collected data for user snoopdogg

UserName	snoopdogg		
Start Date	2008-02-20		
End Date	2010-08-19		
	All	**NvN**	**NvN/All**
# of tweets	2.535	1.309	0,52
# of dis. tweets	4.779	3.609	0,76
# of words	26.410	18.095	0,69
# of dis. words	4.779	3.609	0,76
# of sign. words	15.263	10.715	0,7
# of dis. sign. words	3.720	2.903	0,78
# of nNoun	8.523	5.888	0,69
# of dist. nNoun	2.655	1.856	0,7
# of verbs	5.511	4.036	0,73
# of dist. verbs	1.063	916	0,86
# of RTs	87	61	0,7

Table B.14. The summary of the collected data for user BarackObama

UserName	BarackObama		
Start Date	2007-04-29		
End Date	2010-08-18		
	All	**NvN**	**NvN/All**
# of tweets	849	707	0,83
# of dis. tweets	3.041	2.830	0,93
# of words	15.178	13.022	0,86
# of dis. words	3.041	2.830	0,93
# of sign. words	8.458	7.398	0,87
# of dis. sign. words	2.435	2.279	0,94
# of nNoun	3.775	3.384	0,9
# of dist. nNoun	1.369	1.149	0,84
# of verbs	3.917	3.357	0,86
# of dist. verbs	995	939	0,94
# of RTs	16	12	0,75

Table B.15. The summary of the collected data for user kevinrose

UserName	kevinrose		
Start Date	2009-01-23		
End Date	2010-08-19		
	All	**NvN**	**NvN/All**
# of tweets	3.173	1.151	0,36
# of dis. tweets	5.775	4.334	0,75
# of words	28.969	16.948	0,59
# of dis. words	5.775	4.334	0,75
# of sign. words	16.064	9.687	0,6
# of dis. sign. words	4.614	3.617	0,78
# of nNoun	6.481	4.301	0,66
# of dist. nNoun	2.782	1.943	0,7
# of verbs	7.461	4.306	0,58
# of dist. verbs	1.783	1.438	0,81
# of RTs	119	82	0,69

Table B.16. The summary of the collected data for user TIME

UserName	TIME		
Start Date	2009-07-14		
End Date	2010-08-18		
	All	**NvN**	**NvN/All**
# of tweets	3.054	1.241	0,41
# of dis. tweets	8.477	5.077	0,6
# of words	28.294	13.179	0,47
# of dis. words	8.477	5.077	0,6
# of sign. words	15.101	7.667	0,51
# of dis. sign. words	5.511	3.704	0,67
# of nNoun	6.889	3.921	0,57
# of dist. nNoun	3.323	2.020	0,61
# of verbs	6.837	3.206	0,47
# of dist. verbs	2.181	1.459	0,67
# of RTs	18	16	0,89

Table B.17. The summary of the collected data for user aplusk

UserName	aplusk		
Start Date	2009-07-20		
End Date	2010-08-19		
	All	**NvN**	**NvN/All**
# of tweets	3.159	1.298	0,41
# of dis. tweets	7.004	4.694	0,67
# of words	34.232	19.244	0,56
# of dis. words	7.004	4.694	0,67
# of sign. words	17.705	10.523	0,59
# of dis. sign. words	5.328	3.881	0,73
# of nNoun	7.389	4.874	0,66
# of dist. nNoun	3.137	2.061	0,66
# of verbs	8.316	4.646	0,56
# of dist. verbs	2.070	1.545	0,75
# of RTs	274	167	0,61

Table B.18. The summary of the collected data for user andersoncooper

UserName	andersoncooper		
Start Date	2009-08-25		
End Date	2010-08-19		
	All	**NvN**	**NvN/All**
# of tweets	3.176	1.449	0,46
# of dis. tweets	7.828	4.884	0,62
# of words	29.034	17.007	0,59
# of dis. words	7.828	4.884	0,62
# of sign. words	17.503	10.388	0,59
# of dis. sign. words	4.806	3.557	0,74
# of nNoun	8.439	5.452	0,65
# of dist. nNoun	2.681	1.815	0,68
# of verbs	7.466	4.067	0,54
# of dist. verbs	2.072	1.515	0,73
# of RTs	52	43	0,83

Table B.19. The summary of the collected data for user 50cent

UserName	50cent		
Start Date	2008-12-19		
End Date	2010-08-17		
	All	**NvN**	**NvN/All**
# of tweets	644	316	0,49
# of dis. tweets	1.790	1.270	0,71
# of words	7.830	4.708	0,6
# of dis. words	1.790	1.270	0,71
# of sign. words	4.289	2.739	0,64
# of dis. sign. words	1.306	975	0,75
# of nNoun	2.072	1.395	0,67
# of dist. nNoun	813	555	0,68
# of verbs	1.732	1.067	0,62
# of dist. verbs	477	352	0,74
# of RTs	67	41	0,61

Table B.20. The summary of the collected data for user BreakingNews

UserName	BreakingNews		
Start Date	2010-05-16		
End Date	2010-08-19		
	All	**NvN**	**NvN/All**
# of tweets	3.200	3.025	0,95
# of dis. tweets	10.114	9.835	0,97
# of words	48.258	46.125	0,96
# of dis. words	10.114	9.835	0,97
# of sign. words	33.981	32.572	0,96
# of dis. sign. words	7.876	7.734	0,98
# of nNoun	16.424	15.964	0,97
# of dist. nNoun	4.853	4.395	0,91
# of verbs	15.033	14.255	0,95
# of dist. verbs	2.940	2.884	0,98
# of RTs	0	0	0

Table B.21. The summary of the collected data for user TheOnion

UserName	TheOnion		
Start Date	2009-05-14		
End Date	2010-08-18		
	All	**NvN**	**NvN/All**
# of tweets	3.146	2.120	0,67
# of dis. tweets	10.639	8.494	0,8
# of words	31.413	22.996	0,73
# of dis. words	10.639	8.494	0,8
# of sign. words	20.070	15.120	0,75
# of dis. sign. words	7.491	6.308	0,84
# of nNoun	9.261	7.478	0,81
# of dist. nNoun	4.372	3.375	0,77
# of verbs	8.998	6.453	0,72
# of dist. verbs	2.970	2.458	0,83
# of RTs	14	8	0,57

Table B.22. The summary of the collected data for user cnnbrk

UserName	cnnbrk		
Start Date	2009-12-13		
End Date	2010-08-19		
	All	**NvN**	**NvN/All**
# of tweets	3.069	2.480	0,81
# of dis. tweets	8.752	7.812	0,89
# of words	37.045	30.948	0,84
# of dis. words	8.752	7.812	0,89
# of sign. words	24.753	20.901	0,84
# of dis. sign. words	6.108	5.628	0,92
# of nNoun	11.887	10.445	0,88
# of dist. nNoun	3.598	3.029	0,84
# of verbs	11.161	9.111	0,82
# of dist. verbs	2.452	2.252	0,92
# of RTs	19	17	0,89

Table B.23. The summary of the collected data for user TechCrunch

UserName	TechCrunch		
Start Date	2010-04-19		
End Date	2010-08-19		
	All	**NvN**	**NvN/All**
# of tweets	3.196	2.336	0,73
# of dis. tweets	9.998	8.211	0,82
# of words	36.850	27.841	0,76
# of dis. words	9.998	8.211	0,82
# of sign. words	21.432	16.817	0,78
# of dis. sign. words	6.614	5.675	0,86
# of nNoun	10.607	8.707	0,82
# of dist. nNoun	4.113	3.277	0,8
# of verbs	9.067	6.905	0,76
# of dist. verbs	2.433	2.088	0,86
# of RTs	5	5	1

Table B.24. The summary of the collected data for user dsearls

UserName	dsearls		
Start Date	2006-10-21		
End Date	2010-08-17		
	All	**NvN**	**NvN/All**
# of tweets	2.636	1.396	0,53
# of dis. tweets	7.643	5.617	0,73
# of words	34.006	22.289	0,66
# of dis. words	7.643	5.617	0,73
# of sign. words	17.924	12.156	0,68
# of dis. sign. words	6.093	4.839	0,79
# of nNoun	7.801	5.606	0,72
# of dist. nNoun	3.736	2.682	0,72
# of verbs	8.111	5.274	0,65
# of dist. verbs	2.248	1.793	0,8
# of RTs	72	52	0,72

Table B.25. The summary of the collected data for user davewiner

UserName	davewiner		
Start Date	2010-04-02		
End Date	2010-08-19		
	All	**NvN**	**NvN/All**
# of tweets	2.888	1.290	0,45
# of dis. tweets	7.131	5.047	0,71
# of words	33.152	19.156	0,58
# of dis. words	7.131	5.047	0,71
# of sign. words	16.889	10.157	0,6
# of dis. sign. words	5.521	4.107	0,74
# of nNoun	6.829	4.593	0,67
# of dist. nNoun	3.140	2.112	0,67
# of verbs	8.155	4.553	0,56
# of dist. verbs	2.241	1.704	0,76
# of RTs	0	0	0

Table B.26. The summary of the collected data for user zappos

UserName	zappos		
Start Date	2008-03-08		
End Date	2010-08-19		
	All	**NvN**	**NvN/All**
# of tweets	2.190	1.720	0,79
# of dis. tweets	6.394	5.668	0,89
# of words	39.085	32.260	0,83
# of dis. words	6.394	5.668	0,89
# of sign. words	22.230	18.768	0,84
# of dis. sign. words	5.503	4.950	0,9
# of nNoun	8.907	7.831	0,88
# of dist. nNoun	2.997	2.453	0,82
# of verbs	10.813	8.995	0,83
# of dist. verbs	2.330	2.104	0,9
# of RTs	10	9	0,9

Table B.27. The summary of the collected data for user postsecret

UserName	postsecret		
Start Date	2009-02-11		
End Date	2010-08-18		
	All	**NvN**	**NvN/All**
# of tweets	976	782	0,8
# of dis. tweets	3.725	3.390	0,91
# of words	16.092	13.554	0,84
# of dis. words	3.725	3.390	0,91
# of sign. words	8.662	7.443	0,86
# of dis. sign. words	3.179	2.921	0,92
# of nNoun	3.779	3.349	0,89
# of dist. nNoun	1.700	1.397	0,82
# of verbs	3.910	3.323	0,85
# of dist. verbs	1.413	1.296	0,92
# of RTs	115	68	0,59

Table B.28. The summary of the collected data for user TheEconomist

UserName	TheEconomist		
Start Date	2010-02-14		
End Date	2010-08-19		
	All	**NvN**	**NvN/All**
# of tweets	2.235	1.783	0,8
# of dis. tweets	7.844	6.932	0,88
# of words	29.858	25.481	0,85
# of dis. words	7.844	6.932	0,88
# of sign. words	17.772	15.284	0,86
# of dis. sign. words	5.573	5.077	0,91
# of nNoun	9.401	8.056	0,86
# of dist. nNoun	3.193	2.528	0,79
# of verbs	6.402	5.643	0,88
# of dist. verbs	2.233	2.085	0,93
# of RTs	4	2	0,5

Table B.29. The summary of the collected data for user google

UserName	google		
Start Date	2009-02-26		
End Date	2010-08-18		
	All	**NvN**	**NvN/All**
# of tweets	1.677	1.130	0,67
# of dis. tweets	5.977	4.824	0,81
# of words	22.003	15.980	0,73
# of dis. words	5.977	4.824	0,81
# of sign. words	13.038	9.739	0,75
# of dis. sign. words	4.118	3.506	0,85
# of nNoun	5.581	4.527	0,81
# of dist. nNoun	2.591	1.978	0,76
# of verbs	6.302	4.453	0,71
# of dist. verbs	1.529	1.311	0,86
# of RTs	245	169	0,69

Table B.30. The summary of the collected data for user zef

UserName	zef		
Start Date	2008-03-09		
End Date	2010-08-18		
	All	**NvN**	**NvN/All**
# of tweets	3.109	1.418	0,46
# of dis. tweets	6.883	4.879	0,71
# of words	33.037	19.758	0,6
# of dis. words	6.883	4.879	0,71
# of sign. words	18.285	11.343	0,62
# of dis. sign. words	5.174	3.936	0,76
# of nNoun	7.924	5.283	0,67
# of dist. nNoun	3.070	2.126	0,69
# of verbs	8.140	4.852	0,6
# of dist. verbs	1.919	1.487	0,77
# of RTs	311	169	0,54

Table B.31. The summary of the collected data for user nytimes

UserName	nytimes		
Start Date	2010-06-05		
End Date	2010-08-19		
	All	**NvN**	**NvN/All**
# of tweets	3.198	2.018	0,63
# of dis. tweets	9.668	7.008	0,72
# of words	29.749	19.752	0,66
# of dis. words	9.668	7.008	0,72
# of sign. words	19.230	13.175	0,69
# of dis. sign. words	6.404	4.942	0,77
# of nNoun	9.074	6.812	0,75
# of dist. nNoun	3.719	2.619	0,7
# of verbs	8.758	5.587	0,64
# of dist. verbs	2.683	2.044	0,76
# of RTs	7	6	0,86

Table B.32. The summary of the collected data for user britneyspears

UserName	britneyspears		
Start Date	2008-10-10		
End Date	2010-08-18		
	All	**NvN**	**NvN/All**
# of tweets	425	290	0,68
# of dis. tweets	1.516	1.276	0,84
# of words	6.481	4.983	0,77
# of dis. words	1.516	1.276	0,84
# of sign. words	3.301	2.595	0,79
# of dis. sign. words	1.201	1.026	0,85
# of nNoun	1.648	1.246	0,76
# of dist. nNoun	652	486	0,75
# of verbs	1.304	1.075	0,82
# of dist. verbs	518	453	0,87
# of RTs	27	15	0,56

Table B.33. The summary of the collected data for user jayrosen_nyu

UserName	jayrosen_nyu		
Start Date	2009-11-26		
End Date	2010-08-19		
	All	**NvN**	**NvN/All**
# of tweets	3.170	2.445	0,77
# of dis. tweets	9.452	8.339	0,88
# of words	59.981	47.235	0,79
# of dis. words	9.452	8.339	0,88
# of sign. words	29.420	23.906	0,81
# of dis. sign. words	7.174	6.483	0,9
# of nNoun	11.689	10.219	0,87
# of dist. nNoun	3.903	3.263	0,84
# of verbs	14.498	11.212	0,77
# of dist. verbs	3.000	2.665	0,89
# of RTs	0	0	0

Table B.34. The summary of the collected data for user ijustine

UserName	ijustine		
Start Date	2009-05-16		
End Date	2010-08-19		
	All	**NvN**	**NvN/All**
# of tweets	3.172	1.399	0,44
# of dis. tweets	5.167	3.672	0,71
# of words	38.767	21.875	0,56
# of dis. words	5.167	3.672	0,71
# of sign. words	19.871	11.647	0,59
# of dis. sign. words	4.081	3.089	0,76
# of nNoun	7.132	4.725	0,66
# of dist. nNoun	2.254	1.535	0,68
# of verbs	9.749	5.348	0,55
# of dist. verbs	1.762	1.342	0,76
# of RTs	9	2	0,22

Table B.35. The summary of the collected data for user PerezHilton

UserName	PerezHilton		
Start Date	2010-05-21		
End Date	2010-08-18		
	All	**NvN**	**NvN/All**
# of tweets	3.165	914	0,29
# of dis. tweets	4.859	3.366	0,69
# of words	25.092	13.026	0,52
# of dis. words	4.859	3.366	0,69
# of sign. words	13.404	7.112	0,53
# of dis. sign. words	3.860	2.779	0,72
# of nNoun	6.098	3.333	0,55
# of dist. nNoun	2.252	1.426	0,63
# of verbs	5.666	2.957	0,52
# of dist. verbs	1.460	1.116	0,76
# of RTs	1	1	1

Table B.36. The summary of the collected data for user BryanAlexander

UserName	BryanAlexander		
Start Date	2009-08-02		
End Date	2010-08-19		
	All	**NvN**	**NvN/All**
# of tweets	3.071	2.181	0,71
# of dis. tweets	8.517	7.290	0,86
# of words	43.600	33.191	0,76
# of dis. words	8.517	7.290	0,86
# of sign. words	25.939	20.301	0,78
# of dis. sign. words	7.452	6.463	0,87
# of nNoun	11.077	9.193	0,83
# of dist. nNoun	4.350	3.423	0,79
# of verbs	11.642	8.762	0,75
# of dist. verbs	2.877	2.496	0,87
# of RTs	162	132	0,81

Table B.37. The summary of the collected data for user Scobleizer

UserName	Scobleizer		
Start Date	2010-05-31		
End Date	2010-08-19		
	All	**NvN**	**NvN/All**
# of tweets	2.149	1.357	0,63
# of dis. tweets	4.687	3.890	0,83
# of words	34.993	24.056	0,69
# of dis. words	4.687	3.890	0,83
# of sign. words	17.221	12.270	0,71
# of dis. sign. words	4.038	3.393	0,84
# of nNoun	6.459	5.128	0,79
# of dist. nNoun	2.115	1.618	0,77
# of verbs	8.467	5.677	0,67
# of dist. verbs	1.795	1.493	0,83
# of RTs	10	6	0,6

Table B.38. The summary of the collected data for user smashingmag

UserName	smashingmag		
Start Date	2010-03-23		
End Date	2010-08-19		
	All	**NvN**	**NvN/All**
# of tweets	3.193	1.378	0,43
# of dis. tweets	7.175	4.592	0,64
# of words	31.599	15.936	0,5
# of dis. words	7.175	4.592	0,64
# of sign. words	17.953	9.595	0,53
# of dis. sign. words	4.685	3.307	0,71
# of nNoun	7.613	4.600	0,6
# of dist. nNoun	2.766	1.816	0,66
# of verbs	8.002	3.962	0,5
# of dist. verbs	1.717	1.197	0,7
# of RTs	0	0	0

Table B.39. The summary of the collected data for user THE_REAL_SHAQ

UserName	THE_REAL_SHAQ		
Start Date	2008-11-18		
End Date	2010-08-18		
	All	**NvN**	**NvN/All**
# of tweets	2.519	954	0,38
# of dis. tweets	4.486	3.394	0,76
# of words	21.382	13.374	0,63
# of dis. words	4.486	3.394	0,76
# of sign. words	12.737	8.299	0,65
# of dis. sign. words	4.035	3.091	0,77
# of nNoun	7.043	4.586	0,65
# of dist. nNoun	2.728	1.853	0,68
# of verbs	4.577	3.105	0,68
# of dist. verbs	1.294	1.075	0,83
# of RTs	54	34	0,63

Table B.40. The summary of the collected data for user NiemanLab

UserName	NiemanLab		
Start Date	2009-06-23		
End Date	2010-08-19		
	All	**NvN**	**NvN/All**
# of tweets	3.157	2.420	0,77
# of dis. tweets	10.149	8.856	0,87
# of words	47.366	37.682	0,8
# of dis. words	10.149	8.856	0,87
# of sign. words	26.945	22.150	0,82
# of dis. sign. words	7.174	6.424	0,9
# of nNoun	11.837	10.291	0,87
# of dist. nNoun	4.160	3.428	0,82
# of verbs	11.852	9.333	0,79
# of dist. verbs	2.810	2.527	0,9
# of RTs	33	26	0,79

Table B.41. The summary of the collected data for user uskudarli

UserName	uskudarli		
Start Date	2008-05-07		
End Date	2010-08-24		
	All	**NvN**	**NvN/All**
# of tweets	141	48	0,34
# of dis. tweets	756	435	0,58
# of words	1.421	695	0,49
# of dis. words	756	435	0,58
# of sign. words	817	407	0,5
# of dis. sign. words	596	337	0,57
# of nNoun	340	193	0,57
# of dist. nNoun	325	167	0,51
# of verbs	363	175	0,48
# of dist. verbs	264	147	0,56
# of RTs	24	16	0,67

Table B.42. The summary of the collected data for user leolaporte

UserName	leolaporte		
Start Date	2008-02-27		
End Date	2010-08-18		
	All	**NvN**	**NvN/All**
# of tweets	3.156	1.970	0,62
# of dis. tweets	9.082	7.369	0,81
# of words	47.523	34.896	0,73
# of dis. words	9.082	7.369	0,81
# of sign. words	25.412	19.046	0,75
# of dis. sign. words	7.510	6.397	0,85
# of nNoun	11.092	8.616	0,78
# of dist. nNoun	4.660	3.607	0,77
# of verbs	11.435	8.426	0,74
# of dist. verbs	2.739	2.370	0,87
# of RTs	84	55	0,65

Table B.43. The summary of the collected data for user johnbreslin

UserName	johnbreslin		
Start Date	2009-05-18		
End Date	2010-08-18		
	All	**NvN**	**NvN/All**
# of tweets	2.670	1.490	0,56
# of dis. tweets	7.609	6.041	0,79
# of words	33.157	22.438	0,68
# of dis. words	7.609	6.041	0,79
# of sign. words	18.548	13.133	0,71
# of dis. sign. words	6.030	4.974	0,82
# of nNoun	8.561	6.481	0,76
# of dist. nNoun	3.862	2.891	0,75
# of verbs	7.747	5.231	0,68
# of dist. verbs	2.112	1.757	0,83
# of RTs	377	252	0,67

Table B.44. The summary of the collected data for user Veronica

UserName	Veronica		
Start Date	2009-06-26		
End Date	2010-08-19		
	All	**NvN**	**NvN/All**
# of tweets	3.149	928	0,29
# of dis. tweets	5.271	3.489	0,66
# of words	31.371	14.481	0,46
# of dis. words	5.271	3.489	0,66
# of sign. words	14.895	7.323	0,49
# of dis. sign. words	4.602	3.062	0,67
# of nNoun	5.268	3.009	0,57
# of dist. nNoun	2.480	1.498	0,6
# of verbs	7.210	3.349	0,46
# of dist. verbs	1.962	1.315	0,67
# of RTs	29	15	0,52

Table B.45. The summary of the collected data for user EelcoVisser

UserName	EelcoVisser		
Start Date	2009-04-13		
End Date	2010-08-18		
	All	**NvN**	**NvN/All**
# of tweets	1.250	762	0,61
# of dis. tweets	4.109	3.345	0,81
# of words	16.523	11.815	0,72
# of dis. words	4.109	3.345	0,81
# of sign. words	9.190	6.845	0,74
# of dis. sign. words	3.497	2.910	0,83
# of nNoun	4.145	3.261	0,79
# of dist. nNoun	1.930	1.455	0,75
# of verbs	4.093	2.943	0,72
# of dist. verbs	1.434	1.197	0,83
# of RTs	167	111	0,66

Table B.46. The summary of the collected data for user iamdiddy

UserName	iamdiddy		
Start Date	2010-03-28		
End Date	2010-08-19		
	All	**NvN**	**NvN/All**
# of tweets	2.756	1.603	0,58
# of dis. tweets	4.862	4.024	0,83
# of words	37.694	26.613	0,71
# of dis. words	4.862	4.024	0,83
# of sign. words	19.768	14.571	0,74
# of dis. sign. words	4.086	3.466	0,85
# of nNoun	9.558	7.359	0,77
# of dist. nNoun	2.654	2.106	0,79
# of verbs	8.336	5.937	0,71
# of dist. verbs	1.358	1.165	0,86
# of RTs	799	476	0,6

Table B.47. The summary of the collected data for user Newsweek

UserName	Newsweek		
Start Date	2009-09-16		
End Date	2010-08-19		
	All	**NvN**	**NvN/All**
# of tweets	3.058	1.821	0,6
# of dis. tweets	9.288	6.634	0,71
# of words	33.494	22.539	0,67
# of dis. words	9.288	6.634	0,71
# of sign. words	18.784	12.940	0,69
# of dis. sign. words	6.445	4.967	0,77
# of nNoun	9.482	6.757	0,71
# of dist. nNoun	3.966	2.711	0,68
# of verbs	7.658	5.292	0,69
# of dist. verbs	2.421	1.907	0,79
# of RTs	105	68	0,65

Table B.48. The summary of the collected data for user CBSNews

UserName	CBSNews		
Start Date	2010-04-07		
End Date	2010-08-19		
	All	**NvN**	**NvN/All**
# of tweets	3.020	2.340	0,77
# of dis. tweets	10.469	9.411	0,9
# of words	38.582	32.814	0,85
# of dis. words	10.469	9.411	0,9
# of sign. words	25.858	22.351	0,86
# of dis. sign. words	7.844	7.278	0,93
# of nNoun	12.309	10.946	0,89
# of dist. nNoun	4.576	3.829	0,84
# of verbs	11.728	9.916	0,85
# of dist. verbs	3.207	2.977	0,93
# of RTs	25	21	0,84

Table B.49. The summary of the collected data for user Poynter

UserName	Poynter		
Start Date	2010-03-01		
End Date	2010-08-18		
	All	**NvN**	**NvN/All**
# of tweets	3.197	2.974	0,93
# of dis. tweets	11.821	11.385	0,96
# of words	50.844	47.594	0,94
# of dis. words	11.821	11.385	0,96
# of sign. words	32.095	30.373	0,95
# of dis. sign. words	8.513	8.267	0,97
# of nNoun	16.557	15.900	0,96
# of dist. nNoun	5.368	4.813	0,9
# of verbs	13.055	12.210	0,94
# of dist. verbs	3.025	2.946	0,97
# of RTs	79	68	0,86

Table B.50. The summary of the collected data for user TeamMalachiae

UserName	TeamMalachiae		
Start Date	2010-05-02		
End Date	2010-09-09		
	All	**NvN**	**NvN/All**
# of tweets	1.038	393	0,38
# of dis. tweets	1.927	1.374	0,71
# of words	11.885	5.937	0,5
# of dis. words	1.927	1.374	0,71
# of sign. words	5.660	3.032	0,54
# of dis. sign. words	1.632	1.155	0,71
# of nNoun	2.007	1.289	0,64
# of dist. nNoun	817	544	0,67
# of verbs	2.832	1.354	0,48
# of dist. verbs	749	522	0,7
# of RTs	15	5	0,33

Table B.51. The summary of the collected data for user mathrabbit1

UserName	mathrabbit1		
Start Date	2009-04-04		
End Date	2010-09-09		
	All	**NvN**	**NvN/All**
# of tweets	292	160	0,55
# of dis. tweets	1.264	937	0,74
# of words	3.492	2.256	0,65
# of dis. words	1.264	937	0,74
# of sign. words	1.833	1.205	0,66
# of dis. sign. words	914	685	0,75
# of nNoun	934	606	0,65
# of dist. nNoun	477	306	0,64
# of verbs	746	506	0,68
# of dist. verbs	419	315	0,75
# of RTs	186	120	0,65

Table B.52. The summary of the collected data for user greatspeaking

UserName	greatspeaking		
Start Date	2010-04-15		
End Date	2010-09-09		
	All	**NvN**	**NvN/All**
# of tweets	3.200	2.115	0,66
# of dis. tweets	10.680	8.185	0,77
# of words	32.314	22.394	0,69
# of dis. words	10.680	8.185	0,77
# of sign. words	20.115	14.234	0,71
# of dis. sign. words	7.253	5.853	0,81
# of nNoun	10.947	8.044	0,73
# of dist. nNoun	4.650	3.424	0,74
# of verbs	7.721	5.349	0,69
# of dist. verbs	2.560	2.105	0,82
# of RTs	2	2	1

Table B.53. The summary of the collected data for user stjohnk5

UserName	stjohnk5		
Start Date	2010-08-14		
End Date	2010-09-09		
	All	**NvN**	**NvN/All**
# of tweets	3.199	2.115	0,66
# of dis. tweets	7.777	6.821	0,88
# of words	43.258	33.840	0,78
# of dis. words	7.777	6.821	0,88
# of sign. words	25.313	20.147	0,8
# of dis. sign. words	6.809	5.944	0,87
# of nNoun	14.028	11.254	0,8
# of dist. nNoun	4.743	3.821	0,81
# of verbs	8.946	7.234	0,81
# of dist. verbs	1.943	1.806	0,93
# of RTs	73	50	0,68

Table B.54. The summary of the collected data for user orangeunicorns

UserName	orangeunicorns		
Start Date	2009-10-26		
End Date	2010-09-09		
	All	**NvN**	**NvN/All**
# of tweets	1.548	611	0,39
# of dis. tweets	3.533	2.489	0,7
# of words	14.147	8.116	0,57
# of dis. words	3.533	2.489	0,7
# of sign. words	7.653	4.473	0,58
# of dis. sign. words	3.234	2.266	0,7
# of nNoun	3.585	2.182	0,61
# of dist. nNoun	2.206	1.424	0,65
# of verbs	3.178	1.844	0,58
# of dist. verbs	943	716	0,76
# of RTs	23	12	0,52

Table B.55. The summary of the collected data for user anna0974

UserName	anna0974		
Start Date	2010-04-21		
End Date	2010-09-09		
	All	**NvN**	**NvN/All**
# of tweets	369	150	0,41
# of dis. tweets	1.000	702	0,7
# of words	3.843	2.046	0,53
# of dis. words	1.000	702	0,7
# of sign. words	2.106	1.121	0,53
# of dis. sign. words	824	559	0,68
# of nNoun	817	494	0,6
# of dist. nNoun	394	257	0,65
# of verbs	973	492	0,51
# of dist. verbs	393	251	0,64
# of RTs	0	0	0

Table B.56. The summary of the collected data for user iTauqeer

UserName	iTauqeer		
Start Date	2009-09-26		
End Date	2010-09-09		
	All	**NvN**	**NvN/All**
# of tweets	586	464	0,79
# of dis. tweets	2.149	1.875	0,87
# of words	6.860	5.687	0,83
# of dis. words	2.149	1.875	0,87
# of sign. words	4.322	3.686	0,85
# of dis. sign. words	1.546	1.353	0,88
# of nNoun	2.437	2.123	0,87
# of dist. nNoun	857	682	0,8
# of verbs	1.589	1.336	0,84
# of dist. verbs	667	581	0,87
# of RTs	124	98	0,79

Table B.57. The summary of the collected data for user LUKIKA

UserName	LUKIKA		
Start Date	2010-07-25		
End Date	2010-09-09		
	All	**NvN**	**NvN/All**
# of tweets	3.152	1.699	0,54
# of dis. tweets	7.404	5.526	0,75
# of words	42.794	24.997	0,58
# of dis. words	7.404	5.526	0,75
# of sign. words	20.811	12.994	0,62
# of dis. sign. words	6.155	4.756	0,77
# of nNoun	8.543	6.068	0,71
# of dist. nNoun	3.680	2.619	0,71
# of verbs	9.967	5.761	0,58
# of dist. verbs	2.325	1.801	0,77
# of RTs	380	244	0,64

Table B.58. The summary of the collected data for user Alyssafelldown

UserName	Alyssafelldown		
Start Date	2010-04-02		
End Date	2010-09-09		
	All	**NvN**	**NvN/All**
# of tweets	3.062	587	0,19
# of dis. tweets	3.370	1.754	0,52
# of words	19.110	6.531	0,34
# of dis. words	3.370	1.754	0,52
# of sign. words	10.084	3.565	0,35
# of dis. sign. words	3.138	1.576	0,5
# of nNoun	4.556	1.778	0,39
# of dist. nNoun	1.971	896	0,45
# of verbs	4.323	1.466	0,34
# of dist. verbs	1.098	585	0,53
# of RTs	22	1	0,05

Table B.59. The summary of the collected data for user Hajji_love

UserName	Hajji_love		
Start Date	2008-12-29		
End Date	2010-09-09		
	All	**NvN**	**NvN/All**
# of tweets	935	489	0,52
# of dis. tweets	2.749	2.136	0,78
# of words	12.324	7.927	0,64
# of dis. words	2.749	2.136	0,78
# of sign. words	6.442	4.236	0,66
# of dis. sign. words	2.374	1.862	0,78
# of nNoun	2.627	1.788	0,68
# of dist. nNoun	1.280	898	0,7
# of verbs	2.895	1.892	0,65
# of dist. verbs	987	780	0,79
# of RTs	1	0	0

Table B.60. The summary of the collected data for user CallaLove

UserName	CallaLove		
Start Date	2010-07-21		
End Date	2010-09-09		
	All	**NvN**	**NvN/All**
# of tweets	393	104	0,26
# of dis. tweets	1.121	681	0,61
# of words	3.590	1.460	0,41
# of dis. words	1.121	681	0,61
# of sign. words	1.768	768	0,43
# of dis. sign. words	951	541	0,57
# of nNoun	610	318	0,52
# of dist. nNoun	492	250	0,51
# of verbs	885	358	0,4
# of dist. verbs	434	244	0,56
# of RTs	0	0	0

APPENDIX C: TWITTER KEYWORD SEARCH DATA

Table C.1. The summary of the collected data for keyword Moscow Metro

Keyword	moscow metro		
Start Date	2010-03-29		
End Date	2010-04-02		
	All	**NvN**	**NvN/All**
# of tweets	2.745	2.543	0,93
# of dis. tweets	3.563	3.332	0,94
# of words	50.174	47.988	0,96
# of dis. words	3.563	3.332	0,94
# of sign. words	27.930	26.619	0,95
# of dis. sign. words	2.187	2.078	0,95
# of nNoun	19.155	18.324	0,96
# of dist. nNoun	1.359	1.163	0,86
# of verbs	7.825	7.427	0,95
# of dist. verbs	783	759	0,97
# of RTs	967	943	0,98

Table C.2. The summary of the collected data for keyword spain

Keyword	spain		
Start Date	2010-07-11		
End Date	2010-07-12		
	All	**NvN**	**NvN/All**
# of tweets	191.953	134.489	0,7
# of dis. tweets	102.814	86.162	0,84
# of words	2.382.128	1.955.527	0,82
# of dis. words	102.814	86.162	0,84
# of sign. words	1.465.183	1.208.240	0,82
# of dis. sign. words	73.355	61.437	0,84
# of nNoun	864.445	690.772	0,8
# of dist. nNoun	64.688	51.964	0,8
# of verbs	497.963	435.199	0,87
# of dist. verbs	8.168	7.891	0,97
# of RTs	35.683	26.226	0,73

Table C.3. The summary of the collected data for keyword germany

Keyword	germany		
Start Date	2010-07-01		
End Date	2010-07-09		
	All	**NvN**	**NvN/All**
# of tweets	277.766	185.974	0,67
# of dis. tweets	152.446	124.647	0,82
# of words	3.463.092	2.750.429	0,79
# of dis. words	152.446	124.647	0,82
# of sign. words	2.123.998	1.697.760	0,8
# of dis. sign. words	100.223	82.709	0,83
# of nNoun	1.298.639	994.054	0,77
# of dist. nNoun	89.836	71.486	0,8
# of verbs	672.079	584.684	0,87
# of dist. verbs	9.692	9.334	0,96
# of RTs	37.515	25.029	0,67

Table C.4. The summary of the collected data for keyword oilspill

Keyword	oilspill		
Start Date	2010-05-30		
End Date	2010-05-31		
	All	**NvN**	**NvN/All**
# of tweets	8.460	7.260	0,86
# of dis. tweets	11.165	10.270	0,92
# of words	136.444	121.084	0,89
# of dis. words	11.165	10.270	0,92
# of sign. words	82.560	74.126	0,9
# of dis. sign. words	8.628	8.073	0,94
# of nNoun	44.173	39.248	0,89
# of dist. nNoun	5.467	4.654	0,85
# of verbs	32.674	29.917	0,92
# of dist. verbs	2.973	2.855	0,96
# of RTs	3.729	3.269	0,88

Table C.5. The summary of the collected data for keyword messi

Keyword	messi		
Start Date	2010-07-01		
End Date	2010-07-08		
	All	**NvN**	**NvN/All**
# of tweets	36.905	25.674	0,7
# of dis. tweets	31.862	25.715	0,81
# of words	508.063	399.913	0,79
# of dis. words	31.862	25.715	0,81
# of sign. words	317.495	249.451	0,79
# of dis. sign. words	26.885	21.881	0,81
# of nNoun	206.873	155.455	0,75
# of dist. nNoun	22.208	16.856	0,76
# of verbs	87.860	76.237	0,87
# of dist. verbs	4.390	4.199	0,96
# of RTs	9.013	6.111	0,68

Table C.6. The summary of the collected data for keyword israel

Keyword	israel		
Start Date	2010-05-30		
End Date	2010-06-07		
	All	**NvN**	**NvN/All**
# of tweets	246.539	209.983	0,85
# of dis. tweets	129.861	115.220	0,89
# of words	3.803.254	3.414.489	0,9
# of dis. words	129.861	115.220	0,89
# of sign. words	2.271.390	2.058.199	0,91
# of dis. sign. words	68.931	62.437	0,91
# of nNoun	1.244.323	1.117.052	0,9
# of dist. nNoun	57.316	49.892	0,87
# of verbs	852.568	787.807	0,92
# of dist. verbs	10.668	10.366	0,97
# of RTs	86.471	76.992	0,89

Table C.7. The summary of the collected data for keyword argentina

Keyword	argentina		
Start Date	2010-07-01		
End Date	2010-07-09		
	All	**NvN**	**NvN/All**
# of tweets	127.429	86.759	0,68
# of dis. tweets	84.601	68.213	0,81
# of words	1.609.491	1.263.400	0,78
# of dis. words	84.601	68.213	0,81
# of sign. words	997.290	782.737	0,78
# of dis. sign. words	60.699	49.363	0,81
# of nNoun	616.296	457.762	0,74
# of dist. nNoun	53.006	40.993	0,77
# of verbs	313.731	272.314	0,87
# of dist. verbs	7.264	6.996	0,96
# of RTs	17.621	11.527	0,65

Table C.8. The summary of the collected data for keyword flotilla

Keyword	flotilla		
Start Date	2010-05-29		
End Date	2010-06-07		
	All	**NvN**	**NvN/All**
# of tweets	205.776	179.606	0,87
# of dis. tweets	92.855	83.559	0,9
# of words	3.198.814	2.888.261	0,9
# of dis. words	92.855	83.559	0,9
# of sign. words	1.940.161	1.770.103	0,91
# of dis. sign. words	46.121	43.029	0,93
# of nNoun	1.138.503	1.025.612	0,9
# of dist. nNoun	36.678	32.747	0,89
# of verbs	668.731	625.786	0,94
# of dist. verbs	8.760	8.564	0,98
# of RTs	97.429	86.896	0,89

Table C.9. The summary of the collected data for keyword netherlands

Keyword	netherlands		
Start Date	2010-07-11		
End Date	2010-07-12		
	All	**NvN**	**NvN/All**
# of tweets	77.120	56.207	0,73
# of dis. tweets	43.436	36.634	0,84
# of words	983.899	813.101	0,83
# of dis. words	43.436	36.634	0,84
# of sign. words	618.815	515.550	0,83
# of dis. sign. words	31.100	26.134	0,84
# of nNoun	378.860	306.676	0,81
# of dist. nNoun	25.490	20.076	0,79
# of verbs	199.258	175.768	0,88
# of dist. verbs	5.330	5.061	0,95
# of RTs	14.111	10.377	0,74

Table C.10. The summary of the collected data for keyword brazil

Keyword	brazil		
Start Date	2010-07-01		
End Date	2010-07-09		
	All	**NvN**	**NvN/All**
# of tweets	194.417	136.460	0,7
# of dis. tweets	106.996	89.670	0,84
# of words	2.618.828	2.072.571	0,79
# of dis. words	106.996	89.670	0,84
# of sign. words	1.512.455	1.213.460	0,8
# of dis. sign. words	78.361	66.190	0,84
# of nNoun	863.189	674.243	0,78
# of dist. nNoun	69.313	56.347	0,81
# of verbs	542.561	456.765	0,84
# of dist. verbs	8.542	8.227	0,96
# of RTs	29.297	20.879	0,71

Table C.11. The summary of the collected data for keyword gaza

Keyword	gaza		
Start Date	2010-05-29		
End Date	2010-06-07		
	All	**NvN**	**NvN/All**
# of tweets	163.312	144.573	0,89
# of dis. tweets	85.968	77.711	0,9
# of words	2.542.077	2.324.224	0,91
# of dis. words	85.968	77.711	0,9
# of sign. words	1.576.438	1.456.888	0,92
# of dis. sign. words	40.057	37.138	0,93
# of nNoun	901.185	824.043	0,91
# of dist. nNoun	31.794	28.085	0,88
# of verbs	574.634	542.201	0,94
# of dist. verbs	7.756	7.561	0,97
# of RTs	69.731	62.739	0,9

REFERENCES

1. Facebook, "A social networking website", http://www.facebook.com, May 2010.

2. Myspace, "A social networking website", http://www.myspace.com, January 2010.

3. Wikipedia, "A Free Online Encyclopedia", http://www.wikipedia.org/, 2010.

4. Twitter, "A very popular microblogging service", http://www.twitter.com, April 2010.

5. XML, "Extensible Markup Language", http://www.w3.org/XML/, 1998.

6. RDF, "Resource Description Framework", http://www.w3.org/RDF/, 2004.

7. SPARQL, "SPARQL Protocol and RDF Query Language", http://www.w3.org/TR/rdf-sparql-query/, 2008.

8. Tumblr, "A short form blog system", http://www.tumblr.com, June 2010.

9. Jaiku, "A microblogging system", http://www.jaiku.com, June 2010.

10. South by Southwest, "A private company based in Texas that is known for Music and Media Conference and Festival", http://sxsw.com, May 2010.

11. Chirp, "The official Twitter Developer Conference", http://chirp.twitter.com/index.html, April 2010.

12. CNN, "CNN on Twitter", http://twitter.com/cnn, 2010.

13. BBC, "BBC on Twitter", http://twitter.com/bbc, 2010.

14. NYTimes, "NYTimes on Twitter", http://twitter.com/nytimes, 2010.

15. Twitter API, "Twitter Application Programming Interface wiki", http://apiwiki.twitter.com, September 2009.

16. Twitter4J, "A Twitter API for JAVA Programmming Language", http://twitter4j.org/en/index.html, 2010.

17. MacFadyen, H., "The Parts of Speech in English Language", 2010, http://www.writingcentre.uottawa.ca/hypergrammar/partsp.html.

18. WordNet, "A lexical database for the English language", http://wordnet.princeton.edu/, 2010.

19. Java, A., X. Song, T. Finin, and B. Tseng, "Why we twitter: understanding microblogging usage and communities", *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pp. 56–65, ACM, New York, NY, USA, 2007.

20. Krishnamurthy, B., P. Gill, and M. Arlitt, "A few chirps about twitter", *WOSP '08: Proceedings of the first workshop on Online social networks*, pp. 19–24, ACM, New York, NY, USA, 2008.

21. Milian, M., "Twitter sees earth-shaking activity during SoCal quake.", http://latimesblogs.latimes.com/technology/2008/07/twitter-earthqu.html, July 2008.

22. Cohen., N., "Twitter on the barricades: Six lessons learned.", http://www.nytimes.com/2009/06/21/weekinreview/21cohenweb.html, June 2009.

23. Vieweg, S., A. L. Hughes, K. Starbird, and L. Palen, "Microblogging during two natural hazards events: what twitter may contribute to situational awareness.", Mynatt, E. D., D. Schoner, G. Fitzpatrick, S. E. Hudson, K. Edwards, and T. Rodden (editors), *CHI*, pp. 1079–1088, ACM, 2010.

24. Shamma, D. A., L. Kennedy, and E. F. Churchill, "Tweet the debates: understanding community annotation of uncollected sources", *WSM '09: Proceedings of the first SIGMM workshop on Social media*, pp. 3–10, ACM, New York, NY, USA, 2009.

25. Kwak, H., C. Lee, H. Park, and S. B. Moon, "What is Twitter, a social network or a news media?", Rappa, M., P. Jones, J. Freire, and S. Chakrabarti (editors), *WWW*, pp. 591–600, ACM, 2010.

26. Phelan, O., K. McCarthy, and B. Smyth, "Using twitter to recommend real-time topical news.", Bergman, L. D., A. Tuzhilin, R. D. Burke, A. Felfernig, and L. Schmidt-Thieme (editors), *RecSys*, pp. 385–388, ACM, 2009.

27. Sankaranarayanan, J., H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling, "TwitterStand: news in tweets.", Wolfson, O., D. Agrawal, and C.-T. Lu (editors), *GIS*, pp. 42–51, ACM, 2009.

28. Degirmencioglu, E. A., *Exploring Area-Specific Microblogger Social Networks*, Master's thesis, Bogazici University, 2010.

29. Yurtsever, E., *Sweettweet : A Semantic Analysis For Microblogging Environments*, Master's thesis, Bogazici University, 2010.

30. Akman, D. S., *Revealing Microblogger Interests By Analyzing Contributions*, Master's thesis, Bogazici University, 2010.

31. Strapparava, C. and R. Mihalcea, "SemEval-2007 Task 14: Affective Text", *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pp. 70–74, Association for Computational Linguistics, Prague, Czech Republic, June 2007.

32. ACE-08, "Annotation Tasks", `http://projects.ldc.upenn.edu/ace/annotation/`.

33. WeFollow, "A directory of Twitter users organized by interests", http://wefollow.com, April 2010.

34. Aslan, O., "Web page for MS thesis: AN ANALYSIS OF NEWS ON MICROBLOGGING SYSTEMS", http://elele.cmpe.boun.edu.tr/okaythesis, 2010.

35. Brahaj, A., "A List of English Stop Words", http://armandbrahaj.blog.al/2009/04/14/list-of-english-stop-words/, 2010.