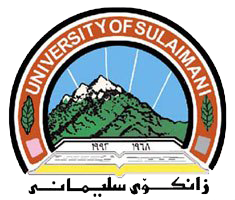
**Social Media Text Preprocessing: Enhancing Data for Sentiment Analysis and Topic Modeling on Twitter**

****

University of Sulaimani

Faculty of Science and Science Education - School of Science

Department of Computer

Data Science

By: Hawraz Abdalla Abubakr

Supervisor

Dr. Miran Taha

Master study; Year 2023-2024

**Abstract**

Twitter is a place where people share quick updates. It's like a big library with lots of up-to-date information. We can use it to understand how people feel (sentiment analysis) and figure out what they're talking about (topic modeling). But, because the words people use on social media can be a bit messy, we need to tidy them up to get useful information. This report explains how we make Twitter data better using different tricks and tools to analyze feelings and discover hot topics.

**Introduction**

This report talks about making Twitter information better so we can understand people's feelings (sentiment analysis) and find out what topics are being talked about (topic modeling). Twitter has a lot of words and symbols that can make it hard for computers to understand. We want to fix this so we can get useful information.

**Why do we want to make Twitter data better?**

1. **Find Out What People Feel:**

**What It Means:** We want to know if people are happy, sad, or something else when they write on Twitter.

**Why It's Important:** Knowing how people feel helps us understand what's important to them.

1. **Discover What's Trending:**

**What It Means:** We want to find out what topics or things are popular or important at the moment.

**Why It's Important:** Understanding trends helps us stay updated on what's happening in the world.

**The Problems We Face**

1. **Tricky Words and Short Texts**

**The Issue:** People on Twitter use short words and may not write everything correctly.

**Why It's a Problem:** Computers can get confused, and we might not get the right information.

1. **Funny Symbols and Pictures**

**The Issue:** Twitter has emojis (small pictures) and symbols that can be hard to understand.

**Why It's a Problem:** Emojis show feelings, and symbols have meanings. We need to make sure we understand them.

1. **Tags and Talking to Others**

**The Issue:** People talk to each other using '@' symbols, and they use hashtags like #fun or #happy.

**Why It's a Problem:** We need to figure out if someone is talking to another person or if a hashtag is about a specific topic.

**How We Make Things Better**

1. **Breaking Down the Words (Tokenization):**

**What It Means:** We break down the words so the computer knows each word separately.

**Why It Helps:** This helps the computer understand what each word means.

1. **Fixing Tricky Words (Normalization):**

**What It Means:** We make all the letters in the words the same (lowercase) so the computer doesn't get confused.

**Why It Helps:** This makes sure the computer doesn't think 'happy' and 'Happy' are different.

1. **Removing Extra Things (Cleaning Up):**

**What It Means:** We take away things like web links and symbols that don't help us know what people are feeling or talking about.

**Why It Helps:** This makes the information clearer and easier for the computer to understand.

**Making Computers Understand Feelings**

1. **Finding the Important Words (Sentiment Analysis):**

**What It Means:** We look at the words to see if they show happiness, sadness, or other feelings.

**Why It Helps:** This helps us know if a tweet is happy or sad.

**Finding Trendy Topics**

1. **Seeing What's Being Talked About (Topic Modeling):**

**What It Means:** We group similar words to find out what topics people are interested in.

**Why It Helps:** This helps us know what's popular or important.

**Conclusion**

Making Twitter data better helps us understand people's feelings and find out what's trending. By fixing words, handling symbols, and using computer tricks, we can get useful information from Twitter. This way, we can learn about what matters to people and stay updated on what's happening around us.

**References**

Cielen Davy Meysman Arno D. B and Ali Mohamed. 2016. Introducing Data Science : Big Data Machine Learning and More Using Python Tools. Shelter Island NY: Manning.

Russell, M. A. (2013). Mining the social web: data mining Facebook, Twitter, LinkedIn, Google+, GitHub, and more. " O'Reilly Media, Inc.".